ACADEMIA
Accelerating the world's research.

The Measure of Madness: Philosophy of Mind, Cognitive Neuroscience, and Delusional Thought

Philip Gerrans

Related papers

Download a PDF Pack of the best related papers

Delusional Attitudes and Default Thinking
Philip Gerrans

Pathologies of hyperfamiliarity in dreams, delusions and déjà vu
Philip Gerrans

Phenomenology and delusions: Who put the 'alien' in alien control?
Tim Bayne

THE MEASURE OF MADNESS

Philosophy of Mind and Cognitive Neuropsychiatry

***

Philip Gerrans

TABLE OF CONTENTS

Acknowledgements

Introduction.

1. The Measure of Madness
   - 1.1. Integrative versus Autonomous Theoretical Explanation
   - 1.2. The Basis of Theoretical Autonomy
   - 1.3. Jaspers and the Inscrutability of Delusion
   - 1.4. Meaning Rationalism and Framework Propositions
   - 1.5. Neurobiological Eliminativism
   - 1.6 Conclusion

2. Models Mechanism and Cognitive Theories
   - 2.1. Cognitive Autonomy: Models and Multiple Realisability.
   - 2.2. Causal Relevance and the Personal Level
   - 2.3. Cognitive Neuropsychiatry and Neurocognitive Psychiatry
   - 2.4 Autonomy Revisited
   - 2.5. The Cognitive Economy
   - 2.6. Theoretical Definition

3 The Processing Hierarchy and the Salience System
   - 3.1. The Processing Hierarchy
   - 3.2. A Computational Framework
   - 3.3. The Salience System and Reward Prediction
   - 3.4. Salience and the Adaptive Critic
   - 3.5. Dopamine and Delusion
   - 3.6 Applications

4. The Default Mode Network
   - 4.1. Simulations as Narrative Elements
   - 4.2. Mental Time Travel and the Default Network
   - 4.3. Delusions as a “Mixed Mode” of the Default Network
   - 4.4 The First Person Perspective and Decontextualisation
   - 2.5. The Default Network and the “Essential Indexical”
   - 4.6. Subjectivity, Affective processing and the Hub of the Default Network
   - 4.7 Anticorrelation Between Default and Decontextualised Processes
   - 4.8 A Mundane Example

5. Dreaming Default Thinking and Delusion
   - 5.1. Dreaming and the Default Mode Network
   - 5 .2. The AIM Model
   - 5.3. Binding, Cognitive Coherence and the Fregoli Delusion

5.4 Feature Binding and the Fregoli Delusion
5.5. Dorsolateral deactivation in dreams and delusions
5.6. Are Delusions Dreams?

6. The Second Factor. Default or doxastic incorporation
6.1. Doxastic Theories and the Second Factor
6.2. Performance Accounts. Endorsement, Explanation and Incorporation
6.3. Interactionism, Explanationism and Attributional style
6.4. Attributional Style and the Cotard delusion
6.5. Bias and Default Thinking
6.6 Competence Accounts: Deficits of Belief Fixation

7. Imagination Incorporated
7.1. Incorporating Imagination
7.2. Belief and Imagination: Congruence and Incongruence
6.3. Joint Incorporation
6.4. The Metacognitive Account
6.5. Delusions and Default processing
6.6. The Dual Nature of Default Thoughts
6.7. Imaginative Resistance and the Essential Indexical
7.8. Cognitive Therapy for Doxastic Theorists
7.9. Imagination and Psychological Structure

8. The Sense of Agency Lost and Found. Experience and Thought in Schizophrenic Delusion.
8.1. The Sense of Agency. Lost and Found
8.2. The Priory of Visual Experience
8.3. Predictive Mechanisms and Cognitive Architecture
8.4. Experimental Evidence
8.5. Awareness of Predictions in Schizophrenia
8.6. Passivity and Externality
8.7. Passivity of Thought
8.8. External Attribution of Thoughts
8.9. External Attribution and Psychological Coherence
8.10. Passivity of Experience, Externality of thought.

9. Louis Sass and the Schizophrenic Lifeworld
9.1. Schreber's Lifeworld
9.2. Cognitive Phenomenology

Bibliography
Index

INTRODUCTION

In 1979 three neuropsychologists reported an unusual case of delusion developed by a man following his admission to hospital in a coma following a car accident. Prior to his accident he had exhibited signs of psychiatric illness such as paranoid delusions and auditory hallucinations. However after his car accident, which produced extensive brain damage (including a necrotic right frontal lobe, a detail which will become important in ensuing chapters) he produced a new series of delusions.

Released from hospital after ten months he returned home to visit his family for the weekend. After this visit,

> He stated that he now lived with a second, different family, virtually identical with his "first" family, and that they lived in a house just like the one he had lived in previously…
>
> He insisted that he had two families of identical composition. The wives of both families had the same given name, the same maiden name, very similar appearance and manner, came from the same town, and had brothers with the same names. There were five children in each family, with the same names and the same sex distribution, a1though he believed the children of his original family were about 1 year younger than those in the second family…He described positive feelings toward "both wives," showed no anger or distress about his first wife's desertion, and specifically expressed thankfulness that she had located a substitute. (Alexander Stuss and Benson 1979, p334)

Alexander Stuss and Benson described this delusion, slightly misleadingly, as a delusion of "reduplication", forming part of family of delusions of misidentification, which typically arise following right hemisphere damage. Over the last two decades philosophers and cognitive scientists have become increasingly interested in these

delusions because their explanation interweaves philosophical questions about the nature of belief and rationality with the explanation of clinical symptoms by cognitive science and neurobiology (Stein, 1996).

In this book I provide an account of what is going in this type of case that extends to other cases of delusion and irrational belief. I argue that we can explain delusion without necessarily invoking a failure or malfunction in a process dedicated to the rational evaluation of beliefs for truth or empirical adequacy. There is a sense in which delusions, and I will argue, many thoughts, *bypass* processes of rational belief fixation. If this is correct then the explanation of delusions, and many other species of irrational belief, does not require us to treat them as beliefs produced by reasoning deficits or biases. Rather we need to explain how it is that delusions can arise in response to experience, be maintained and interact with other mental states to control behavior, without necessarily being *believed*[1](Dennett, 1991).

Of course for many philosophers and psychologists the fact that delusions are psychologically embedded in this way automatically qualifies them as beliefs. For these theorists someone with a delusion of misidentification who murders their father while deluded that they are attacking an imposter surely *believes* their delusion.

I am not so sure. Delusions certainly resemble straightforward empirical beliefs: they are formed in response to perceptual information and often regulate behavior. Equally, however, they have some properties that make them difficult to assimilate to straightforward cases of empirical belief. They are often extremely *subjective,* in the sense, not only of being produced as a response to personal experience, but somehow insulated from public standards of justification, openness to dialogue and disconfirmation by obvious counterevidence. Similarly, when pressed, delusional patients often show a

[1] Using belief in what Rorty called the "industrial strength" sense which implies a mental state *constitutively* responsive to evidence and rational norms. In a much weaker sense, for example the way Dennett uses the term, as an interpretive heuristic for predicting behavior, I suppose delusions might qualify. But such usages don't gain any traction on the cognitive phenomena and, in Davidsonian hands, lead to autonomy theses (see chapter 1 and 2). Part of the aim of this book is to clear a path for the explanation of phenomena such as delusions without being drawn into debates about the nature of belief.

lack of commitment to their delusion they do not show for other empirical beliefs. There is also something about the phenomenology of delusion, the first person experience of delusional life, that seems to distinguish it from other modes of experience that lead straightforwardly to empirical belief.

Of course the fact that a mental state is based on strange and intractable forms of experience, is subjective, maintained with ambivalence, and insulated from disconfirmation according to intersubjective standards does not show it is not a belief. Delusions might be beliefs that arise and are maintained in unusual ways. Much of the research on delusion takes place from within this *doxastic* framework, trying to explain how it is that the neural and cognitive correlates of delusion can introduce things like confirmation biases, dysfunctional gaps, deficits, or shortcuts into a belief-fixation system. As with the heuristics and biases tradition in psychology, the idealisations of logic and probability theory provide a standard of competence against which irrational human performance is initially measured and explained. For example the differing performance of delusional and non-delusional people on an inductive reasoning task has been used to explain the difference between delusional and non-delusional belief fixation (Garety and Freeman, 1999, Garety et al., 1991, Hemsley and Garety, 1986, Huq et al., 1988, Startup et al., 2008, Broome et al., 2007).

My argument in this book is that we can set aside questions about the nature of rationality and still satisfactorily explain the puzzling belief-like properties of delusion, as well as its puzzling non-belief-like properties. The methodology I recommend is close attention to, and description of, the cognitive properties of the neural mechanisms that produce delusion. Thus the approach I take is consistent with that proposed by Dominic Murphy in his plea for a (cognitive neuro) scientific psychiatry: “we arrive at a comprehensive set of facts about how the mind works, and then ask which of its products and breakdowns matter for our various projects” (Murphy, 2006, p. 105).

If our project is to explain why someone says that their family has been replaced by doubles it is not clear that the concept of belief, anchored as it is in conceptual debates about the nature of rationality, needs to play a large role in that explanation. Rather to

explain delusion we describe the components of the information processing system active in delusion and trace the flow of information through the system, making use of the evidence available from different disciplines ranging from molecular neurobiology to clinical psychiatry and cognitive neuroscience.

A clue to the cognitive nature of delusion is provided by the patient himself in an interview with his doctors (my italics).

E. Isn't that [two families] unusual?
S. It was *unbelievable*!
E. How do you account for it?
S. I don't know. I try to understand it myself, and it
was virtually impossible.
E. What if I told you I don't believe it?
S. That's perfectly understandable. In fact, when I tell the story, *I feel that I'm concocting a story*. . . It's not quite right. Something is wrong.
E. If someone told you the story, what would you think?
S. I would find it *extremely hard to believe*. I should be defending myself.

I think that the patient is right. He is telling a story. The story is consistent with his experience (though not with his knowledge about the nature of reality) but from the point of view of a third person it is unbelievable. Nonetheless he cannot abandon that story, or revise it to fit with his wider empirical knowledge, and continues to act according to it. Hence his ambivalence when pressed about its empirical adequacy when treated as a description of reality.

I think that all of these phenomenological and psychological facts can be accounted for in terms of the cognitive processes that produce the delusion. And that *those processes do not include a system dedicated to the fixation of belief according to standards of consistency and empirical adequacy*. The most important cognitive system involved is the default mode network, (DMN). The DMN represents personally relevant information as narrative elements. Such narratives elements are not always assembled into full-scale

narratives, they may remain fragmentary and episodic, but their cognitive nature is to be the building block of a story assembled from subjective experiences. Consequently any representation assembled solely from such narrative elements can only be *subjectively adequate.*

I call such building blocks "default thoughts". They are produced by the DMN, which was named to reflect the automatic tendency of the mind to revert to this type of self-referential processing. This tendency is manifest in a variety of active and passive conditions from personal deliberation and planning (mental time travel as it is known to cognitive neuroscience) to daydreaming, dreaming and delusion. Each of these conditions represents a particular mode of default processing. These modes are characterised, not only by the intrinsic properties of the DMN, but by the way it communicates with other cognitive systems in each condition.

One reason that delusions are naturally treated as empirical beliefs is that their linguistic expression does not signal their cognitive nature. The same thought such as "my wife has been replaced by an imposter" or "I am under surveillance by the CIA" can occur in the mind as narrative fragment triggered by experience or as a factual description of reality. Because sentences or thoughts can perform both roles (being part of a subjectively adequate story or part of an empirically adequate description or explanation) the mere fact of their utterance or thought doesn't determine which role they are playing. Nor, I shall argue does the fact that people, including delusional people, *act on the basis* of such thoughts.

Which role the thought is playing depends not so much on whether the subject *acts* on it, or how she intends it (it is quite possible that someone intends to be expressing empirical beliefs but is actually telling a subjectively adequate story) but whether the mind is configured to treat the thought as a narrative element, description or part of a causal explanation. In the latter cases the mind is configured to enable a default thought to be tested for empirical adequacy and consistency with other descriptive or theoretically-anchored beliefs. However I shall argue that in delusion this is not the case.

The argument is not a purely conceptual one or based entirely on phenomenological facts such as the clinical interview above. It integrates evidence from different disciplines about the way the mind configures itself in response to incoming information according to the way neural mechanisms influence cognitive processing. It then applies the resulting account of neural processing and cognitive architecture to case studies of specific delusions of misidentification, alien control and thought insertion in schizophrenia. *En route* I address methodological and conceptual questions about the nature of this type of integrative explanation; the difference between cognitive models and cognitive theories; and the relationship between explanations of the same phenomenon, (delusion) based in phenomenology (lived experience) neural and cognitive theory. The overall idea is that Murphy is right. Our best understanding of the mind comes from understanding cognitive architecture. However that argument cannot be established *a priori* for *all* mental phenomena. The best we can do is construct, revise and, ultimately, unify case-by-case explanations.

In chapter one I argue that collecting and collating correlations between neural, phenomenological and cognitive properties of the delusional mind is useful but we need a theoretical approach that fits all this information together. I set the scene for that approach by describing, and rejecting, two versions of the autonomy thesis: the idea that neurobiological and phenomenological explanations are theoretically insulated from each other. Versions of this autonomy thesis come from neuroscience and from philosophy and phenomenology and have a venerable history, dating from asylum psychiatry of the nineteenth century. I consider modern versions (Neurobiological Eliminativism and Meaning Rationalism) and argue that they rest on implausible notions of the aim and scope of psychiatric explanation. It is normal practice to explain phenomena such as amnesia or macular degeneration in terms of the way neural circuits implement the cognitive processes involved in memory or perception. This suggests that the way to explain psychology and phenomenology in terms of neurobiology is via a cognitive theory. Why should the same approach founder in the case of delusion? I reject the idea that the approach cannot work for the explanation of so-called “personal level”

phenomena, those whose characterization involves normatively-contrained concepts of rationality and agency.

Chapter two develops this idea and considers the form such a cognitive theory should take. It first considers an influential objection to the project of cognitive theorizing, often raised by cognitive scientists themselves. Namely, the argument from multiple - realisability. This argument says that because that cognitive models abstract from details of physical implementation they cannot be used as guides to the nature of actual human performance. A model of visual processing can be implemented in a laptop, camera *or* brain. It may simulate human performance, in the sense of mapping inputs to outputs, without necessarily describing actual human cognitive architecture. When, however, such a models describe actual human cognition they function as cognitive *theories* of specific domains. Cognitive theories are sensitive to the way an actual system implements a computational process. A theory of human vision, for example, explains how the human visual system constructs representations on the basis of retinal input. In order to produce such theories we need a way of describing the phenomenon (e.g. vision, memory, delusion) in such as way that theorists working on different aspects of the phenomenon, ranging from the molecular to the phenomenological and even social, can see how they are all talking about the same thing. This description I call a theoretical definition.

With a (provisional and revisable) theoretical definition at hand, integrative causal explanation proceeds by way of intervention and manipulation. This framework for explanation of complex multilevel phenomena has been articulated by Carl Craver for the case of neuroscience. He emphasises that in order to intervene in a system to produce predictable effects (e.g. to explain how chemical processes produce long term potentiation in neurons which encode memories) we need an adequate description of the mechanisms involved. In the case of delusion these mechanisms are a complex hierarchy of information processing systems implemented in neural wetware. Information flows in the hierarchy according to principles of error detection and correction. Delusions arise at the highest levels of the hierarchy when information becomes salient in the form of default thoughts.

The theoretical definition of delusion I propose to explicate this process is

Delusions arise when ***default cognitive processing, unsupervised by decontextualized processing*** **is** ***monopolised*** **by** ***hypersalient information.***

None of the key terms is defined in a vocabulary proprietary to particular levels of explanation (philosophical, psychological, cognitive, neurobiological, molecular) and it is possible to demonstrate via the interventionist methodology how each term is realised at levels studied by different disciplines in delusional and non-delusional minds. In fact in many cases the nature of a neural process is only fully explained when the cognitive process it is implementing is explained. The salience system is an especially good example. It turns out that the behaviour of dopamine molecules at the synaptic level is best understood when they are treated as part of a mechanism that allocates processing resources to relevant information. Furthermore, the properties of the dopamine salience system would not have been as well understood without the evidence gained in the pharmacological treatment of psychosis, including delusion.

Importantly none of the key terms uses a concept like belief or rationality whose normative status has been claimed by philosophers and autonomy theorists from different disciplines to apply uniquely to the "personal level" of rational agency, at which delusion presents as irrational belief. I argue however that delusion can be explained entirely in a non-normative conceptual framework, which applies to all levels of investigation. For example salience, the functional concept that refers to allocation of processing resources to relevant information, has analogues at all levels of the mind. Reasoning and belief do not. Nonetheless my theoretical definition adequately explains, without invoking the concept of irrational belief or faulty inference, why the patient thought his family had been replaced by doubles.

Chapter three explicates the nature of the cognitive hierarchy and explains each of the key terms. It gives examples of the way information flows through the hierarchy. The

basic concept is error correction. Errors, detected by systems specialised for particular cognitive tasks, are referred to higher-level supervisory systems, which provide feedback to correct those errors and cancel error signals. The highest levels of supervision are default processing and decontextualised processing, which together confer the capacity for stimulus-independent reflection and deliberation. The salience system operates at all levels in the hierarchy optimizing the allocation of cognitive resources to relevant information.

Chapter four explains the nature of default processing, its neural substrate, cognitive properties and role in the cognitive hierarchy. It evolved to produce narrative fragments that enable mental time travel: the simulation of episodes of a subject's life to enable planning, deliberation and reflection. The representations produced by the default system are essentially indexical. They provide a first personal perspective on information in virtue of their dependence on circuitry involved in motivational and emotional processes. The more objective, third person, perspective on information required for establishing truth or consistency of beliefs, requires the activation of decontextualised processing systems. In a well-calibrated mind decontextualised processing *supervises* default processing allowing subjective narratives to be revised to fit reality. In a delusional mind this level of supervision is unavailable or compromised, leaving the subject at the mercy of her default system.

Chapter five explains the nature of default processing in more detail, concentrating on its outputs in different modes: dreaming, mental time travel and delusion. It shows how the phenomenology in each condition is sensitive to neurobiological processes. It concludes that delusion is a mixed mode of cognition, sharing some of the properties of each of the modes. This mixture explains its belief-like and non-belief-like properties. It includes a case study of delusions of misidentification, which have been the focus of a lot of interdisciplinary work in the last three decades.

Chapter six situates this account against more standard doxastic accounts, the one and two factor accounts as they are known. From within a doxastic framework the rationality

or otherwise of delusion is the main *explanandum*. Some doxastic theorists have argued that delusions are produced by "broadly rational" processes of belief fixation responding to anomalous sensory or perceptual information processing (the one factor account). Two factor theorists have argued that, while anomalous perceptual or sensory processing is typically the trigger for delusion, some additional problem (the second factor) in cognitive processing is required to account for the fact that the delusion is produced and maintained in the face of its inconsistency with background beliefs. This problem can be conceptualized as a deficit (competence) or bias (performance) in belief fixation. Two factor accounts are supported by evidence of abnormal functioning in the right dorsolateral prefrontal cortex in delusion. I argue that the evidence from cognitive neuroscience suggests that this neural system is best conceptualized as the substrate of decontextualised processing system that supervises the default network. I also argue that that this description of the relevant cognitive architecture sufficiently explains delusion. There is no need to appeal to normative considerations derived from folk psychological conceptions of rational agency, inference and belief to explain the difference between delusional and non-delusional minds.

Chapter seven addresses an issue for this minimalist cognitive approach. Namely, delusions are default thoughts produced by a simulation system that evolved to enable stimulus-independent thought. As such, default thoughts are essentially imaginative. But the everyday conception of imagination does not treat imagination as a state that causes behavior. We act on the basis of belief not imagination. I argue that in fact we do act on the basis of imagination, more often perhaps than we realise, *and* that the idea of action based on default thought actually explains, not only delusion, but many cases of behavior based on apparently irrational beliefs which are problematic for doxastic accounts.

Chapter eight is a case study of some typical schizophrenic delusions: of alien control and thought insertion. The account I provide is not of schizophrenia *per se* but of two of the delusional symptoms. These delusions are important to the account because they highlight the necessary role of cognitive theory in linking the neurobiological and phenomenological levels of explanation. The phenomenology in question is loss of a

"sense of agency" for voluntary movements, and the neural correlate is abnormal patterns of activity in the parietal cortex. I show that a cognitive theory of visuo-motor control embedded in the overall architecture of cognitive control outlined in earlier chapters is required to explain why high levels of activity in these regions produce loss of a sense of agency. This case study also shows, via a discussion of competing views, that cognitive theorising is highly contentious. The view I propose is a minority one but it requires fewer assumptions and cleaves more closely to the neural and cognitive data than some others. In fact some other accounts have ignored the distinction between models and theories, assuming that a cognitive model which correlates with the neural phenomena is thereby, via some theoretically-loaded assumptions, vindicated as a cognitive theory.

Chapter nine reconsiders the nature of delusion from the phenomenological perspective. Louis Sass is just one theorist who has argued that close attention to phenomenology suggests that doxastic accounts cannot do justice to the way it feels to be delusional. On his view delusions express the result of intense absorption in experience and attention to its felt quality. This explains the preoccupied and introspective character of some schizophrenic delusions. The subjectivity and ambivalence of some delusions, combined with intact reasoning on non-delusional subjects, suggest to Sass that the subject is keeping two sets of mental books. One for engaging in the delusional world whose limits are set by the way things feel or seem, the other, for the intersubjective world, responsive to public norms of empirical belief fixation. Sass's account is very important because it can often seem that approaches to delusion such as those discussed in previous chapters leave experience aside in favour of developing and refining cognitive theories. Clinicians and historians of psychiatry often find such approaches inadequate, precisely because the nature of delusional experience seems to drop out of the picture to be replaced by flow charts and equations. I argue that the flow charts and equations of cognitive theory are important, not because they replace a focus on experience, but because they can help explain experience. In fact the theoretical definition of delusion given in chapter two is entirely consistent with accounts like that of Sass. It explains experience rather than explaining it away.

THE MEASURE OF MADNESS

He was particularly distressed by the scaly appearance of his skin, which he believed was caused by a lizard growing inside his body, the lizard's skin being evident on his arms and legs. He gave the growth of the lizard inside his chest as the reason for stabbing himself. He related this to an incident 10 years before when, in Jamaica, a lizard had run across his face. He believed the lizard had 'left its mark' and that a curse had produced his skin lesions (Browning and Jones, 1988).

Integrative versus autonomous theorising

How should we explain delusion? A constant flood of research provides information about the neural correlates of delusion at levels of resolution ranging from the molecular to the synaptic and neuroanatomical. At the same time cognitive neuroscientists have developed information-processing theories which target the contribution to delusion of cognitive processes such as memory, perception, sensory processing, reward prediction, emotion, inference, attention and executive functioning. Integrating all this research with the evidence gained in clinical encounters and by experimental psychologists is a project whose time is at hand. However many theorists across the relevant disciplines argue that fully integrative explanation is impossible. All that we can hope for is the accumulation of correlations between things like the expression of genes involved in the synthesis of dopamine and glutamate, levels of activity in particular neural circuits and the psychological features of delusion: the abnormalities of experience, belief and behaviour which bring people to clinical attention.

Sad to say the role of philosophy here has been, by accident or design, to reinforce arguments that integrative explanation is impossible. This book takes a different tack. I aim to show that we can explain the contribution to delusions, like the one above, of facts like the distribution of dopamine receptors in the prefrontal cortex, as well as more familiar "personal level" facts about the nature of delusional experience, thought and behaviour . We can show how facts identified and explained by disciplines operating at "levels" such as molecular neurobiology or neuroanatomy can explain psychological and phenomenological level facts that give delusion its clinical profile.

In general people who think that integrative explanation is impossible do not dispute a causal role for neurobiological and cognitive processes in psychiatric disorder. However they argue that such causal explanations, even though they explain the mechanistic functioning of human organisms, cannot work as explanations of delusion (Bermúdez, 2000b, Davies, 2000a, Davies, 2000b, Davies, 1990, Bennett and Hacker, 2003). The reason is that delusions are problems of personhood. Persons are organisms capable of self-awareness, conscious reflection on their own experience and rational control of behavior (Graham, 2009). Yet, so the argument goes, personal level phenomena can only be explained in terms of other personal-level phenomena: beliefs for example are formed when *persons* try and explain their experiences using their background beliefs (Davidson, 1980, McDowell, 1996, Thornton, 2000). Personal level explanations are theoretically *autonomous* on this view. Interestingly this autonomy thesis has been advocated by both neuroscientists working at molecular and synaptic levels and by psychiatrists and philosophers who concentrate exclusively on personal level phenomena. Equally interesting is the fact that these arguments have existed in more or less unchanged form since asylum psychiatrists in the nineteenth century attempted the intellectual project of transforming correlation to explanation for psychiatric disorder.

I shall respectfully try to undermine the autonomy thesis, not by *a priori* refutation of arguments on which it depends but by producing integrative explanations of some of the most intensively studied (at least by philosophers and philosophically minded-cognitive psychologists), delusions over the last two decades. I shall argue that a full understanding

of personal level phenomena such as delusion requires us to understand persons as complex, hierarchically-organised information-processing systems implemented in neural wetware. In fact if we treat personal level explanation as theoretically insulated from other forms of explanation we lose information, not just about the brains and bodies of human organisms but persons themselves. The way to incorporate that information is via suitable *cognitive* (i.e. information processing) theories that bridge the gap between neurobiological and personal level explanation.

Before half the potential readers close the book, on the basis that existential feelings of dread, despair or elation or the complex patterns of (ir)rational thought they initiate cannot be reduced to computational processes implemented by mechanistic functioning of the brain, I need to emphasise that the type of cognitive explanation I propose does not involve ignoring personal experience and thought or explaining it away by redescribing it as something else. My aim is to show that personhood is a cognitive phenomenon constituted by the fact that that personal-level phenomena, such as feelings, beliefs, emotions and desires arise at the highest levels of a cognitive processing hierarchy whose nature can be described and explained. And delusions, like many psychiatric disorders, emphasise the dependence of personal level phenomena on the complex layers of processing which support personhood.

The basis of theoretical autonomy

Commitment to theoretical autonomy follows from the idea that mechanistic and personal explanation have incompatible aims. Personal explanation makes other people intelligible. This notion of intelligibility or meaningfulness has two, inter-related, aspects which tend to be emphasised by psychological and phenomenological accounts respectively. Rational reconstruction of another person's psychological structure and empathetic identification with their experience. Rational reconstruction makes another person intelligible by attributing her with a similar pattern of perceptual experiences, consequent beliefs about the world and emotional responses to it. If we do not share

sufficient psychological structure we cannot understand them. And, we might add on behalf of autonomy theorists, we get no insight into this type of psychological structure by collecting information about correlates of delusion such as genetic deletions, neuroanatomical abnormalities, P300 waveforms or Blood Oxygen Level Dependent signals detected in magnetic resonance scanners. We gain insight by finding out more about another person's beliefs and motives.

For those who emphasise empathetic identification, to understand another person is to be able to imaginatively inhabit their inner world. The two approaches are not of course distinct. Empathy requires some grasp of psychological structure and we can more readily determine psychological structure if we can empathise. And once again it is hard to see how accumulating facts about the brain, described in the language of neuroscience can assist with the project of empathising.

Jaspers and the inscrutability of delusion

For Karl Jaspers who pioneered modern versions of the autonomy thesis the primary problem in the explanation of delusion was unintelligibility. Ultimately he thought that delusions were best understood in terms of the nature of the *phenomenological encounter* between the deluded and their interpreter (Berrios, 1992, Jaspers, 1963). One feature of that encounter is that in extreme cases there does not seem to be enough sharing or overlap in very fundamental attitudes to experience to enable the outsider to share the lifeworld of the delusional patient. Delusion, especially in the form of so-called bizarre delusion, confronts us with a mind whose structure is opaque to us.

Jaspers did not deny that delusions originated in an organic process. However he thought that organic disorder made sympathetic rationalisation of schizophrenic behaviour not only irrelevant but impossible. The schizophrenic "becomes something inaccessible and foreign, which for this very reason the language defines as deranged." (Sass, 1992)

Jaspers pointed out in his *General Psychopathology*, that some delusions, for example that the subject is being followed, or is dying of cancer are *intelligible* even though they may be bizarre. We can identify imaginatively with the subject of such a delusion in two ways. We can empathise, sharing some of the same emotions. One does not have to believe that someone is dying of cancer to sympathise with the distress they feel as a result of a delusion that they are. Equally we can understand related actions and beliefs, which follow from the delusion even if they are objectively irrational. Thus imaginative identification is partly emotional, partly intellectual and depends on the assumption that that, despite her delusion, the subject coherently integrates volition, emotion and belief in the typical way. In other words the subject retains a sufficient degree of the psychological coherence presupposed by everyday psychology.

However there are other delusions for which this degree of empathy and understanding seem impossible. Sass reports Nijinsky saying in the course of a schizophrenic episode "Once I went for a walk and it seemed to me that I saw some blood on the snow. I followed the traces of blood and sensed that somebody *who was still alive had been killed*" (Sass 1992 p.17). Another example might be the delusion that a subject's bodily organs had been surgically removed during the night without leaving a trace. Another might be the case of a subject who while looking at tables in a restaurant became convinced that the world was coming to an end.

In these cases it seems impossible to identify empathetically with the subject of the delusion or to reconstruct her thought processes. These types of delusions put the subject on the other side of what Jaspers described as the "most profound distinction in psychic life between what is meaningful and allows empathy, and what, in its particular way is *ununderstandable*, mad in the literal sense, schizophrenic psychic life (even though there may be no delusions)" (Sass 1992 p17)

Here Jaspers points out that the life of the schizophrenic with bizarre delusions, as lived from the inside, is something to which we have no empathetic access. Another example of this empathetic inaccessibility is provided by accounts of the prodromal phase of

schizophrenia, in which people report feelings that the world and their experience of it has changed in eerily disturbing and significant ways which cannot be articulated.

Jaspers agrees with the asylum psychiatrists (of which he was an eminent one before becoming a philosopher) that the ultimate cause of schizophrenia is organic brain disease, but focuses instead on the characteristic experience. He does not think that the schizophrenic is like an Alzheimers patient: rather something has happened which reconstructs the patient's mind and renders it inaccessible. Empathy requires a degree of psychological coherence and the closely related project of rational interpretation requires that psychological states be integrated according to some principle of rationality, but it is impossible to reconstruct the experience or reasoning of persons with bizarre or primary delusions

On this view what is definitive of delusion is not dementia but *Inscrutability to an interpreter*. Eugene Bleuler said that after decades working in an asylum his patients lives were as alien to him as those of the birds in his garden.

Meaning Rationalism and framework propositions

Meaning Rationalism takes up the challenge of rendering delusions intelligible. The main task for a Meaning Rationalist is to explain how an apparently irrational delusional belief can exist in the network of beliefs that comprise the agent's doxastic psychology. They focus on the question "Why would someone believe *that proposition*, given the rest of the beliefs in their network?" For example "How, given some obvious biological and zoological beliefs, could someone form the belief they are physically inhabited by, or transmogrifying into, a lizard after having a lizard touch their face?" Meaning rationalism involves applying the conceptual framework of radical interpretation and meaning holism to the problem of inscrutability identified by Jaspers (Campbell, 2001a, Bortolotti, 2005, Bortolotti, 2009).

The essential idea of radical interpretation is that we make other people intelligible by attributing them with intentional states that fit together in a coherent pattern. In the case of beliefs we attribute those beliefs that, if part of a consistent set, would, together with the subject's perceptions, emotions and desires explain her actions. In order for other people to be interpretable in this way they must share our experiential world and processes of belief fixation, at least to the extent that we can plausibly extend the same intentional concepts to them (Bortolotti, 2005). Radical psychological incoherence makes people unintelligible. Of course intelligibility is a matter of degree and radical interpretation does not require complete compliance with the norms of rationality (three year olds are largely intelligible even though their grasp of the concept of consistency is partial at best).

If the words and deeds of delusional subjects seem (partially or wholly) inscrutable the Meaning Rationalist has different options. She could preserve the meaning of the words by either

(i) assuming that the deluded subject's words mean the same thing as ours but she is having very different experiences. The difficulty here is with delusions for which we have no precedent in experience. What for example is the experience of being simultaneously alive and dead? Or of being physically inhabited by a lizard?

(ii) assuming that she is rationalising those experiences in different but still consistent ways. However the glaring inconsistencies of some delusions with the rest of the subject's beliefs (Is it scientifically possible to be inhabited by a lizard, or to be alive and dead?) undermine this idea. Equally it is sometimes difficult to reconstruct the reasoning processes of delusional subjects and preserve any resemblance to what meaning rationalists call "canonical" norms of rationality.

(iii) Saying that the delusions are meaningless and the subject is simply unintelligible. The difficulty with this idea is that although delusions may be bizarre, many are not, and even in the case of bizarre delusion it does appear that delusional subjects are using their words with the same meanings. When

> Paul Schreber said that his psychiatrist's soul was lodged in his (Schreber's) body in the form of a cobweb he meant the same things by 'soul' 'body' and 'cobweb' as his clinicians. And the patient quoted at the beginning of the chapter uses 'lizard', 'scales' and 'curse' in the normal way. That is why his condition is so horrible.

Another option is to say that the words of delusional subjects have meaning within *a purely subjective yet internally coherent psychology not subject to public norms of belief fixation*. This however is not really a live option for Meaning Rationalism. The reason is that meaning rationalism assumes inferential role semantics. A proposition gets its meaning from its relationship to the rest of the propositions, empirical and theoretical, believed by the subject. The content of any single belief can only be explained in terms of its role in a network of other beliefs which stand in relations of mutual justification (Fodor and Lepore, 1992, Bortolotti, 2005). Crucially the rules of justification which weave the web of belief are *publicly accepted rules of inference*. So idiosyncratic inferences or rules of inference cannot lead to meaningful beliefs.

Faced with these difficulties recent versions of meaning rationalism have argued that delusional subjects can be made intelligible on the assumption that their delusions are instances of a change in framework propositions.

The notion of a framework proposition originates with Wittgenstein who noted that there are beliefs on which the coherence of persons psychology depend that do not themselves depend on other beliefs. The belief that one has a body or that the sky is above, provide a framework for experience and thought but, at least within everyday life, do not require support from other beliefs. They provide the bedrock for other epistemic encounters in which we exercise processes of belief fixation, but are not themselves hypotheses generated to explain experience. This notion of framework beliefs is an epistemological concept that immunises certain beliefs from doubt because the epistemological project of justification in terms of evidence and inference does not apply to them, at least within everyday contexts. Wittgenstein puts the point by saying that if someone doubted these

foundational propositions one would not think she had made a mistaken inference but was “deranged”.

Or as Robert Klee (Klee, 2004) says “is delusional”. John Campbell (2001) and he offer an explanation of delusion which preserves the intelligibility of delusion by suggesting that the irrationality of delusion consists in the adoption of a different set of framework propositions. The delusional subject holds to her beliefs so tenaciously because they have become framework propositions: bedrock assumptions which are not justified with reference to other beliefs or experiences.

Meaning Rationalism captures the phenomenology of, especially, bizarre delusion. The delusion has a fundamental and central importance in the psychic life of the subject and seems not only immune to evidence or argument but not produced by any rationally reconstructable process of belief fixation.

However there are several difficulties with meaning rationalism as a global account of delusion: the first is that the strangeness of delusion seems to depend on the fact that delusional subject’s grasp of public meaning seems largely intact. The schizophrenic *really means* that someone is inserting thoughts into her mind. The second is the abandonment of the idea that delusions are responses to experience, which also seems to run counter to the evidence in many cases. Delusions of infection by invisible parasites, for example are often responses to the experience of itching or skin lesion (Bayne and Pacherie, 2004). As Broome (Broome, 2004) notes progressively deeper engagement with delusional patients make even apparently bizarre delusional responses more intelligible even if the clinician cannot entirely imaginatively inhabit the delusional world. That this is possible suggests that both that meaning has not changed entirely and that the delusion represents a response to experience not a seismic cognitive readjustment.

Finally meaning rationalism runs very close to kind of anti realist instrumentalism about psychology in which the coherence of a subject’s psychology becomes a matter of

interpretative ingenuity or community standards for rationalisation. This would insulate delusion from cognitive or neurobiological investigation. As with inscrutability accounts, organic damage is hypothesised to play a causal role in introducing the drastic change in psychological structure but plays no explanatory role. A fully-fledged meaning rationalist may be happy with this conclusion but, as the cases we discus in coming chapters make clear, it is an inadequate response to the facts about the origin of many delusions. Perhaps the most obvious case is delusions of misidentification in which a patient will claim that a familiar person has been replaced by an imposter. The evidence is overwhelming that these delusions result from abnormal activity in neural circuitry which implements a cognitive system specialised for the recognition of faces. The outcome at the personal level is that the patient sees a familiar person but does not experience the phenomenology of recognition. The delusion is a response to that mismatch. The precise nature of the mismatch, response and consequent phenomenology are the objects of cognitive theory described in chapters 4 and 5.

At this point we note only that to say that such misfiring processes *cause but do not explain* the delusion seems perverse. A better approach is closer investigation of the nature of face processing and its malfunction in delusion. Especially if the alternative explanation offered is the claim that the person with the delusion now accepts a hinge proposition that his wife has been replaced by an imposter as a result of an organic process whose nature is irrelevant to the content of the delusion. That way of putting things redescribes the problem (the person has an apparently irrational belief he will not abandon) in personal level vocabulary, rather than explaining it.

This type of case suggests that what is required here is more attention, not less, to the neural and cognitive processes involved in the genesis of delusion.

Meaning rationalism derives its autonomy thesis from the top down, by limiting its explanation to the personal level of delusion. Other autonomy thesis are derived from the bottom up.

Neurobiological Eliminativism

As with Meaning Rationalism Neurobiological Eliminativism has its roots in asylum psychiatry of the nineteenth century.

The pioneer of Neurobiological Eliminativism is Emil Kraepelin, the asylum psychiatrist who first described dementia praecox (schizophrenia) and manic-depression (these were two of only three categories of psychiatric illness recognised by asylum psychiatry. The next edition of the DSM will recognise over 400). Kraepelin's innovation was to treat patterns of co-occurring symptoms as indicators of underlying disease of the central nervous system, standing in the same relationship of effect to cause as the symptoms of any disease. For Kraepelin the organic aetiology ruled out treating psychiatric symptoms as the result of a psychological process since the brain disease impairs thought in the same way as other dementing brain diseases. For him schizophrenia is a *neuropathology*.

> in dementia praecox, partial damage to, or destruction of, cells of the cerebral cortex must probably occur, which may be compensated for in some cases, but which mostly brings in its wake a singular, permanent impairment of the inner life GP 1899 (Kraepelin, 1896)

the effect was to render the patient in his famous words "an orchestra without a conductor".

In this sense neuropathological accounts are genuinely eliminativist, proposing that we can eliminate psychological accounts and/or replace them with explanations of the difference between normal and abnormal neurobiology. Nancy Andreasen for example once claimed

> We need not look to theoretical constructs of 'the mind' or to influences from the external environment… Instead we can look *directly at the brain*" Andreasen 1984 (Andreasen and Grounds, 1984).

Andreasen's pursuit of this approach to schizophrenia culminates in her theory of schizophrenia as the result of asynchronous activity in circuitry looping through the cerebellum, thalamus and frontal cortex. Her quarry is an integrated theory of all schizophrenic symptoms, not just delusion and hallucination, and in particular she aims at explaining the apparent distintegration of the schizophrenic mind which renders severe cases an orchestra without a conductor.

Andreasens's approach is first to identify schizophrenic symptoms and map them as far as possible to their local or distributed neural substrates. We then explain their co-occurrence in terms of neurobiological processes.

According to Andreasen co-occurrence of schizophrenic symptoms derives from their dependence on a common neural circuit which runs from the cerebellum to the prefrontal cortex to the thalamus and back (CCTCC). Integrated mental functioning depends on the synchronised activation in circuitry integrated by the CCTCC. Hence she names the malfunction at the basis of schizophrenia *dysmetria*. (Andreasen et al., 1998, Andreasen, 1999)

It is part of the theory that cognitive dysmetria is not localised in the CCTCC (although it originates there) but is produced in more remote areas by malfunctions in the CCTCC in virtue of their projection to other areas. It is, paradigmatically, a misconnection syndrome. Although the CCTCC is the main circuit it affects with other local neural assemblies, not only via neural circuitry but via dopamine pathways whose origin it intersects.

> Patients suffering from schizophrenia have a misconnection syndrome which leads them to make abnormal associations between mental representations and to lack the ability to distinguish between the self and the not self or the important and the trivial.

> The multiple stimuli that bombard our consciousness cannot be normally suppressed, and the important cannot be distinguished from the unimportant. Therefore internal representations may be attributed to the external world, leading to hallucinations. Perceptions or other information will be misconnected with innappropriate associations leading to delusional misinterpretations. On-line monitoring of language or thoughts will be impaired or mistimed , leading to disorganised speech and disorganised thinking. Behaviour will not be adequately monitored, leading to social awkwardness, excessive aggressiveness or shyness, or other abnormalities in behavioural activities. The inability to monitor may also lead to the "freezing" characteristic of catatonic immobility or a failure to inhibit that is expressed as catatonic excitement. Difficulties in inhibiting or prioritising may also lead to the various negative symptoms such as avolition or anhedonia, much as a computer locks when it cannot match signals sent at an incorrect rate or to an incorrect place. ((Andreasen, 1999)

As Kraepelin himself put it if the neuropathology leaves psychological capacities intact we might need to ascend to the level of psychology to explain the symptoms but the essence of neuropathological accounts is that psychological abnormalities are merely diagnostic not constitutive of delusion. Consequently explanation is exhausted by description of neurobiological mechanisms involved rather than appeal to the psychological structure of the disorder.

The obvious problem for such neuropathological accounts is that schizophrenia, except in the extremes of psychosis, does not look like a dementing disorder in which psychological structure has disintegrated entirely.

This difficulty is of course even more pronounced for delusions like misidentification of persons and infestation, which are circumscribed and monothematic. In such cases breakdown of a specific cognitive subsystem combined with the introduction of reasoning bias or malfunction centred on a single topic seems the most promising hypothesis. Of course the neuropathological account may not be intended to cover such

cases, but in that case it cannot be the basis of a global account of delusion. Equally, if the neuropathological account *is* the correct global account then psychological and cognitive theories of circumscribed and monothematic delusion need revision.

It is worth noting just how revisionist this type of eliminativist approach is, even for perceptual or sensory systems. Macular degeneration following diabetes requires explanation at the cellular and molecular level. However describing those changes does not explain the phenomenon unless the role of the affected cells in visual processing (a cognitive process) and visual experience (a personal level phenomenon) is also described and explained. Scotoma or blurring of perception as e*xplananda* make it clear that perception is a representative process: the patient loses the ability to represent the visual scene. Not just at the cognitive level but at the personal level of visual experience. Consequently what needs to be explained here is not just the causal interactions among neurones but the way those interactions enable cognitive processes and experiences. Visual experience, which is a personal level phenomenon, (the visual cortex doesn't see red, the person does) emerges as the result of a complex hierarchy of computational processes performed by neural circuitry.

Of course when all goes well we do not need to pay much attention to the nature of this hierarchy. We could content ourselves with analysing vision at the personal level. But even commonplace phenomena such as astigmatism direct us to the nature of the mechanisms on which visual experience depends. In fact we understand less not more about vision and visual experience if we restrict ourselves to personal level explanation. Why would this not be true of psychiatric disorder?

Another way to put the point is to ask why it is good explanatory practice to incorporate evidence from different levels using information processing theories for disorders of perception, but not disorders of belief?

Andreasen herself gives the game away in her explanation of the nature of the disruptions caused by dysmetria. To take just one example:

internal *representations* may be attributed to the external world, leading to *hallucinations*. Perceptions or other *information will be misconnected* with *innappropriate associations* leading to *delusional misinterpretations*. (Andreasen 1999, p) (my italics)

Thus when she comes to explain the consequences of neural abnormality Andreasen invokes the *cognitive* properties of neural circuits. Some schizophrenic symptoms can only be explained in representational terms. But to verify these representational hypotheses cognitive theories at different levels of resolution are required.

So Andreasen is not, strictly speaking, an eliminativist at all but an advocate for bottom up approaches to psychiatric explanation. Her eliminativism is more a rhetorical strategy in the continuing unproductive battle between absolutist biological psychiatry and those psychiatrists convinced that neurobiological approaches to the mind essentially aim to explain it away. In fact one reason for her advocacy of the disease model is that in an either/or clinical environment in which psychiatric disorders are *either* brain diseases *o*r maladaptive patterns of thought to be understood and remedied in purely personal terms, emphasising the biological basis of disorders is an important corrective. In some ways she aims to perform a similar service to that performed by psychiatrists who emphasised biological bases of autism in opposition to destructive theories that autism is a personal-level psychological response to unempathic parenting.

Once we restore the idea that the neural circuits disrupted in delusional disorders perform cognitive functions, the relationship between psychology, phenomenology and neurobiology can be explained in terms of the effects on cognition of neurobiological processes. Not only that but cognitive theories of a *particular* process can never be

sufficient for the explanation of delusion because delusions arise at the highest levels of cognitive control. So the cognitive hierarchy, the flow of information from level to level, and the role of particular processes in the overall hierarchy needs to be described.

Once this is done many neuropathological or eliminativist accounts look less eliminativist. For example the ideas of another nineteenth century psychiatrist who pioneered the neuropathological approach, John Hughlings Jackson. fit well with the idea that the mind is hierarchy of cognitive processes. In fact Hughlings Jackson had that idea but he lacked a cognitive theory to substantiate it.

For Jackson (1958) the mind was an hierarchy of systems, implemented in specialised brain structures, in which abstract thought and reasoning emerged as the result of inhibition of the basic sensorimotor systems. Psychiatric disorders arose when “lower” structures became overactive, overwhelming their inhibitory masters or when inhibitory processes weakened as a result of disease in the higher cortical structures which subserved them.

> The higher nervous arrangements evolved out of the lower to keep down those lower, just as a government evolved out of a nation controls, as well as directs, that nation. If this be the process of evolution, then the reverse process of dissolution is not only 'a taking off' of the higher, but is at the very same time a 'letting go' of the lower. If the governing body of this country were destroyed suddenly, we should have two causes for lamentation: (1) the loss of services of eminent men; and (2) the anarchy of the now uncontrolled people. The loss of the governing body answers to the dissolution in our patient (the exhaustion of the highest two layers of his highest centers); the anarchy answers to the no longer controlled activity of the next lower level of evolution (third layer). (Jackson, 1958)Selected Writings II, p. 58)

The disorder on which he did most of his work was epilepsy, which he explained as the result of nervous impulses in temporal regions dedicated to sensorimotor activity.

> I have for more than ten years, and before the experiments of Hitzig and Ferrier were made, held that convolutions of contain nervous arrangements representing movements. It is in accordance with this belief that I have long considered chorea, and more lately convulsion, to be movements resulting from 'discharges' of the cerebral cortex. (Selected Writings I, p. 3)

The point is not that some disorders have no organic basis and others do. All psychiatric disorders result from activity (discharge as Jackson called it) in the central nervous system, however not all that activity is a *cognitive* process. Jackson is claiming plausibly that a cognitive explanation is not required for the symptoms of epilepsy. A more controversial, to our ears, claim is that delusions are similarly non-cognitive neurobiological processes. This amounts to saying that the thoughts and deeds of delusional subject stand in the same relationship to their organic causes as the words of epileptics undergoing seizure or the profanities of someone emerging from anaesthesia.

Clearly this is not correct but the reason is that in the conditions described by Jackson the control hierarchy has either disintegrated or is semi-randomly damaged. In delusion it is largely intact although the flow of information through the hierarchy is different in a delusional mind.

It is worth noting that Jackson may not have equated delusional disorders with epilepsy but simply pursued the non-cognitive neuropathological account as the only suitable method available in the absence of any convincing theory of the relationship between psychological states and organic processes. Freud took the opposite horn of this nineteenth century dilemma. He ultimately despaired of a finding theoretical link between organic processes and personal level psychological states and posited unconscious psychological processes with personal properties.

Cognitive Phenomenology

I have discussed two versions of the autonomy thesis because they are different responses to the challenge of linking neurobiological explanations of organismic function to personal level explanation. Their failures are also suggest the form of a solution. Namely that we treat the human brain as an information processing system in which personal level states (such as feelings and beliefs) and processes such as thinking and reasoning emerge at the highest levels of processing.

I don't think it is possible to show that this is possible (or impossible) *a priori.* Rather we need to attempt the project and consider objections as they arise. In the next two chapters we briefly outline the structure of such a theory before applying it to specific delusions.

.

MODELS MECHANISMS AND COGNITIVE THEORIES

Cognitive Autonomy

The previous chapter suggested that despite the arguments of Meaning Rationalists and Neurobiological Eliminativists personal level psychology is not an autonomous explanatory realm. In fact there *must* be an explanatory relationship between neuroscience and folk psychology (as the everyday practice of personal level explanation is sometimes called). To take an obvious case, amnesia following traumatic brain injury drastically compromises personal level psychology. The person with amnesia will experience and react to the world quite differently after her injury. Insofar as her beliefs and ability to deploy them in reasoning depend on memories, she will also think about the world differently. Memory loss caused by brain damage seems the most plausible hypothesis here. An important step in completing our understanding of disorders like amnesia involves producing the correct account of the way the brain encodes information acquired in experience and then reconstructs representations of that information when subsequently cued. This is the type of project I mean by cognitive theorising.

In fact the most serious objection to this approach comes, not from neurobiological eliminativists or meaning rationalists but from cognitive psychologists who regard *cognitive* explanation as theoretically autonomous. If cognitive explanation is a genuinely autonomous realm then it cannot play the bridging role between personal and neurobiological explanation.

The argument for the autonomy of cognitive explanation is correct in one (but fortunately only one) sense. Namely when applied to cognitive *models* considered independently of their physical implementation. For example if we succeed in describing the algorithms involved in parsing sentences or mapping retinal information to representations of the visual scene which caused the retinal array we can run and test them on a laptop. So in principle computational processes can be realised in both brains and laptops (Lewis, 1980). Not only that but the same computational problem problem can be solved in different ways, even in the same machinery as anyone who has switched word-processing programs can attest. Consequently individual differences between human brains, and the computational solutions they have developed, should make us wary of assuming that two people who behave in the same way in response to a stimulus are processing information identically (Gerrans, 2002a). This is especially so for belief, that paradigm personal-level state. The same belief can be the product of testimony, theoretical reasoning or bitter personal experience and in each case the information processing and neural machinery involved will differ. It follows that there is no *necessary connection* between the computational structure of a problem, its algorithmic solution and the nature of the mechanisms which implement it (Coltheart, 2010, Plaut and Patterson, 2010).

This line of thinking leads to the idea that cognitive theorists produce *models* which operate at a level of abstraction which makes it impossible in principle or in practice to draw any conclusions about the tole played in cognition by neural mechanisms. All cognitive theorising can add to our understanding is models that help with the process of correlation but not explanation (Craver, 2001). For example we might learn that faces can be recognised by features or by configuration and that damage to different areas of the brain correlates with impairments to the different types of process. But that takes us no closer to understanding *how* the brain implements feature-based processing and why, therefore, damage to particular cells affects it in different ways (Patterson and Plaut, 2009, Arbib, 2003).

Causal relevance and the personal level

In order to avoid this problem we need a methodology in which cognitive theories play a different role: pointing toward the way human brains actually perform cognitive processes. The type of explanation I have in mind is that advocated by Carl Craver in *Explaining the Brain* (Craver, 2007, Craver, 2002).

Craver's work is a sustained argument that explanation in *neuroscience* involves the identification and description of relations of *causal relevance* between levels[2] of systemic organization (Craver, 2002). Causal relevance is defined in terms of manipulability and intervention (Woodward, 2008, Woodward). For example unless we understand the mechanisms of memory at all levels we cannot understand the effects on memory of interventions at levels as different as phosphorylation within the neuron, calcium channels, temporal lobe lesion, or such paradigmatically personal encounters with the world as such as armed combat or neglectful early childhood (Lupien et al., 2005, Lupien et al., 2007). Events at leach of these levels are causally relevant to the encoding, storage and reconstruction of representations of information about life events. They are causally relevant in the sense that a difference at one level produces a change at another level. This difference-making is a consequence of the fact that these "levels" are actually descriptions of processing components of the same mechanism, the brain, at different levels of resolution. Thus the essence of explanation in neuroscience is description of

[2] The notion of "levels" is ubiquitous but not everyone uses it the same way. It can refer to ordering relationships between theories (physics is a theory at a lower level than physics); the objects of theories ordered by size or complexity (cells are smaller and less complex than the organs they comprise); functional analyses (vision is a higher level property than edge detection); levels of mereological containment (parts are at a lower level than wholes). These uses can overlap.

The notion of level relevant to this book is levels of *mechanism*. Higher levels correspond to the organization of components, lower levels to components. Interlevel explanation explicates the *causal relevance* of events at different levels.

mechanism(s) at a level of detail which explains the dependency relationships between causally relevant variables (Woodward, 2008).

Note that this way of putting things assumes that the personal level of explanation is simply another level of description of the information processing system. I think that this is correct: personal level explanation is a way of explaining systemic functioning understood as the outcome of a complicated hierarchy of cognitive processes. We have experiences, reflect on them and deliberate about possible actions as a result of complicated information processing which includes the representation of our own and other minds in folk psychological terms. Systematic patterns are detectable and explicable at different levels of functional organisation and the personal level is the highest level for human beings. This is why it makes sense to say that persons remember and have amnesia, but component mechanisms of the memory system do not (Graham, 2009). Similarly persons have beliefs, desires and emotions as a result of the way their brains process relevant information. A person's beliefs for example are the way the system represents to itself the information which percolates to the top of the information processing hierarchy in formats which make it available for reflective thought and verbal communication.

Consequently personal level explanation is ineliminable for psychiatry. Another way to put that in the interventionist framework is to say that some of the essential variables involved in multilevel explanation must be personal (Campbell, 2007). Consider the case of memory again. What people feel, believe, attend to, care about and communicate influences what they remember. If I come to *believe* my future depends crucially on my performance in an exam I might over-activate a stress response which interferes with my ability to remember what I have learnt (Arnsten, 1998). Similarly, high levels of stress, experienced as personal feelings of unease or panic, prevent the encoding of new memories because high levels of corticotrophins, released from the brainstem as a response to the detection of adverse events, deactivate prefrontal circuitry required for sustained attention and working memory (high level cognitive processes), ultimately

preventing long term potentiation (LTP) in neural circuitry (Lupien et al., 2007). Thus memory functions differently under stressful conditions. Explanation might go the other way as well. LTP might be adversely affected by very low level factors, such as changes in cell metabolism produced by lack of sleep, operating at molecular levels unmediated by personal level processing. The consequence will, however, be manifest at the personal level as gaps in memory and consequent faulty inferences. A full explanation of these effects which are only manifest in personal encounters with the world will require tracing and retracing the interactions between mechanisms which enable the personal level representation of information (my explicit beliefs and conscious feeling states for example) and the molecular level mechanisms of memory. It is for this reason that Craver identifies explanation in neuroscience in terms of the description of mechanisms which enable *interlevel mutual manipulability*.

Similar considerations apply to the explanation of the experience and patterns of thought which present as failures of rationality such as delusion. My approach differs from Craver's only in the additional emphasis I place on *computational* mechanisms underpinning personal level phenomena. In the case of delusion, and indeed any train of thought, we are dealing, not just with the (re)construction of a particular representation (your memory of your first day at school for example) but its manipulation in combination with other representations as part of a goal-directed train of thought. So explanation requires the description of those cognitive processes that enable reflection and deliberation. This idea is not new. A journal called *Cognitive Neuropsychiatry*, founded in 1995 now exists alongside *Biological Psychiatry* and Tony Stone and Andy Young's (1997) paper *Delusions and Brain Injury* inaugurated the decade of philosophical theorising which inspired this book.

Cognitive Neuropsychiatry and Neurocognitive Psychiatry

The essential idea of cognitive neuropsychiatry is that without a cognitive theory the problem identified by autonomy theorists in chapter 1 cannot be solved. The gap between neurobiology and psychology will be unbridgeable.

> So called "biological psychiatry" has often attempted to move from clinical label to biochemical deficit without paying attention to the *systemic level of analysis that could mediate between behaviour and material substrate* … The danger of course is that this approach will lead to extensive (and expensive) searches for the biological substrate of non-existent entities. (Halligan and Marshall, 1996)

Interestingly some of the main proponents of cognitive neuropsychiatry are also committed to the autonomy thesis for cognitive theorising. This is because as we noted above a *model* of a computational process abstracts from details of physical implementation (it can be a flow chart, equation, sentence, algorithm or any other abstract representation) so it need not describe an actual mechanism (Marr, 1982).

However to play a role in complete explanations, which identify relations of causal relevance, such models need to be supplemented by descriptions of *mechanisms* which realise them. Thus I prefer the term *neurocognitive theory* to *cognitive model*. Cognitive models are abstract depictions of cognitive architecture: the information processing structure of the mind. Neurocognitive theories explain the how these models are implemented in neural circuitry and hence describe mechanisms. These neurocognitive theories in fact provide "the *systemic level of analysis that could mediate between behaviour and material substrate"* required by Marshall and Halligan.

Thus the explanation I am advocating here extends the project of cognitive neuropsychiatry quite considerably. Cognitive neuropsychiatry employs the modelling methodologies of cognitive psychology, which abstract from details of neural implementation (Shallice, 1988). In effect I aim to treat psychiatry as a branch of

cognitive neuroscience by employing cognitive models that do not abstract away from, but are sensitive to, details of neural implementation. Often the same model can be employed in both projects. A cognitive model of motor control can explain visuomotor coordination for a robot *or* human. The same model, *if* it captures important properties of the human visuomotor system can help us understand why disorders of motor control arise under specific conditions of neural functioning. In such a case the model becomes part of a neurocognitive theory of motor control (Blakemore et al., 2002, Frith et al., 2000, Wolpert and Ghahramani, 2000). Such theories help explain why, to take a case which will become relevant in chapter 8, schizophrenics with high levels of activity in the inferior parietal lobule, feel that they are performing actions they did not intend.

To sloganise, what I advocate here is not *cognitive neuropsychiatry* as proposed by autonomy theorists about cognitive modelling. Rather I advocate *neurocognitive psychiatry*. That is the use of neurocognitive theory to explain psychiatric disorder.

I share the basic idea of cognitive neuropsychiatry that the mechanisms involved in psychiatric disorder necessarily include cognitive, that is information processing, ones. This is in fact a general truth about human psychology. Genes build proteins, which build neurones, which build neural circuits which process information in representational formats which allow the organism to control its behaviour. Tracing the psychological effects of a genetic deletion, a neurotransmitter imbalance, a head injury, a childhood trauma, a private school education or a conversation requires us to treat the mind as an integrated set of computational mechanisms implemented in neural wetware.

I follow Dominic Murphy who emphasises that psychological explanation, of which psychiatric explanation is a species, is causal explanation (Murphy, 2006) and therefore needs to be integrated with cognitive neuroscience. I follow Craver in adding that causal explanation in psychology is a matter not of describing causes, which would involved

settling deep metaphysical issues about the ultimate structure of reality, but *explicating causal relevance* analysed in terms of manipulation and intervention.

Autonomy revisited

Murphy notes that psychiatry is bedevilled by controversies, like those described in chapter 1, generated by treating causal and personal level explanation as mutually exclusive. His main point is that autonomy theses are generated by a misunderstanding of the nature of the theoretical bridge between neurobiology and psychology. It is true that one cannot simply *translate* the vocabulary of feeling and thought into the vocabulary of genetics or ion channels. Nor can the bridging role be played by a set of bridge laws linking neural, cognitive and psychological laws. (If Davidsonians are right there are no psychological laws).

However the approach advocated here does not involve translation or reduction via bridge laws. It involves showing how components of a very complex mechanism, the human mind, work, by explicating relationships of casual relevance between what we might call macro and microcomponents (Bunge, 2003).

Macrocomponents are personal states like thoughts and feelings which can only be conceptualised at the level of the system as a whole, interacting with other similar systems: the person. Microcomponents can be extremely micro. The distribution and sensitivity of a subtype of dopamine receptors in the prefrontal cortex has large consequences for high-level cognitive processes and personal states. And this sensitivity is a consequence of molecular structure (Abi-Dargham et al., 2002, Seamans and Yang, 2004, Williams and Goldman-Rakic, 1995, Braver and Cohen, 2000).

So the main task of this book is to identify mechanisms which enable personhood and explain how they interact with other cognitive mechanisms in order to explain some delusions.

The task is partly philosophical, as opposed to purely empirical, for different reasons. The first is that the empirical science in the area is incomplete. Many issues remain about the evaluation of competing arguments and the relationships between theory and evidence. Secondly the idea that personal psychology can be explained by describing mechanisms and their mutual manipulability is itself a philosophical position that needs defending. That defence cannot however be made *a priori* but depends on explanation of particular cases. So the explanation of delusion is a case study in applied philosophy of science which takes psychiatric explanation as its target (Murphy, 2006).

That case study follows and augments the approach suggested by the pioneering work of Stone and Young (1997) who first articulated the philosophical relevance of the cognitive approach to delusion. They divided the cognitive explanation of delusion into stages corresponding to the hypothesised sequence of processes leading to delusion. Problems with peripheral processes (those which control the mind's sensorimotor interface with the world) are hypothesised to form the first stage of delusion formation (Stone and Young, 1997, Davies et al., 2001). A typical "first stage" problem is explaining the origin of the experience of "hearing voices" in schizophrenia. Is hearing voices an auditory hallucination, the equivalent of "seeing things", or is it the experience of "inner speech"? How does evidence obtained in the lab and imaging magnet bear on this question? This is a surprisingly difficult question to answer (Blakemore et al., 2000, David, 2004, Shergill et al., 2003) but it has important consequences for the explanation of the consequent delusion that someone else is inserting thoughts into the patients mind. Similar difficult questions arise in delusions of misidentification. What exactly is involved in misidentifying a familiar person? How could they appear indistinguishable from the familiar but not be experienced as familiar? (Breen et al., 2000, Ellis and de Pauw, 1994, Ellis et al., 1993a, Ellis and Lewis, 2001) What should we make of the fact that these

delusions have neural correlates in circuitry implicated in the identification of faces? (Schweinberger and Burton, 2003). In each case I argue that the question can be explained (at least in principle) by cognitive neuroscience and explore the consequences for explanation of the resultant delusion.

People with delusions do not just have anomalies of peripheral (that is sensory and perceptual) cognition. They are responding to those anomalies in an unusual way. This response is the " second stage" of delusion formation. Problems at the second stage in are characterised in folk psychological terms as " false beliefs produced by incorrect inference". To see how these beliefs arise we need a theoretical explanation of the normal path from peripheral to "central", that is metacognitive, processing involved in detecting and responding to lower level anomalies. Thus explaining delusions involves producing a cognitive account of the mechanisms which enable humans to reflect on and respond to new information.

A crucial problem for this approach is defining the problem in a way that avoids the "expensive and extensive search for non existent entities". Or, as we might put it, allows us to pursue the project of manipulation and intervention at different levels. Often, and this is the loaf of truth at the core of autonomy theses about personal level psychology, definitions drawn from folk psychology cannot meet this criterion since the personal level phenomenon thus defined turns out not to be causally relevant. For example a definition of episodic memory as *retrieval* of stored experiences simply cannot play that kind of explanatory role no matter how faithful it is to the everyday (folk) conception of memory. The neuroscience, computational theory and experimental psychology of episodic memory converge on the idea that each episode of episodic memory is *constructed* not *retrieved* (Buckner et al., 2008, Schacter et al., 2008, Schacter et al., 2007). This is a case where the folk psychological concept of memory needs to be revised to fit the evidence. It would not do to maintain that, whatever science discovers, memory is not a constructive process, or that scientists have changed the subject.

The same considerations apply to any definition of memory that relies on a normative notion like knowledge ("memory is knowledge of the past caused by previous experience"). No one knows how to demonstrate causal relevance of *normative* epistemic concepts to neural events. Indeed many philosophers have argued that such relevance can never be demonstrated (Bermúdez, 2000b, Davies, 2000b, Gerrans, 1999). Persons are rational agents but their brains are not. And although cognitive processes are rule-governed they typically do not represent the rules as normative for themselves: the rules are "wired in", implicitly represented in the physical structure of the mechanism. The idea of normative obligation seems to apply uniquely at the personal level. This is in fact another major argument for the autonomy of the personal level explanation(Searle, 1983).

If it were the case that personhood could *only* be explained in terms of normative concepts such as rationality then that might be an argument in favour of the autonomy thesis for personal level states. After all delusions are characterised normatively as irrational beliefs, that is beliefs which violate *rational norms* of consistency and coherence.

However it is not true that to explain personal level phenomena we must stay at the personal level. In fact the opposite is the case. A good explanation of personhood needs to descend to the lower levels of explanation that invoke *different* explanatory concepts and entities to those used in folk psychology. This is in fact a general truth about complex systems. We do not need to look for analogical or isomorphic structures at lower levels in a system in order to explain the emergence of structure at higher levels in the system.

The Cognitive Economy

Economics provides a good example. Macroeconomic theory provides a formal mathematical way to predict relations between overall or global systemic states of an economy. For example macroeconomic theory includes theorems about the relationship

between savings and interest rates that enable *ceteris paribus* predictions. It does not follow that individual consumers, firms or households compute that function when they decide to put money in the bank (more likely they thought the girl in the building society advertisement was pretty, the besuited financial adviser wise and reliable, or they panicked when thinking of their impoverished old age). Nonetheless the economy *ceteris paribus* marches in tune with macroeconomic theory.

Similarly the mind is a complex aggregate of interacting systems which, when all goes well allow the organism to cope with the world. Consequently, under ideal conditions, the mind marches, *ceteris paribus*, in tune with procedural rationality: it forms true beliefs and acts accordingly. Thus we can use norms of belief fixation to predict and explain other people. It does not follow, however, that the component subsystems of the mind are governed by procedural rationality anymore than the *ceteris paribus* applicability of macroeconomics implies that individual consumers regulate their actions by computing macroeconomic functions.

In psychology, as in economics, explaining the systemic properties of the mind requires attention to subsystems and their interactions. And, as in economics, this is particularly so *when prediction fails*. If a rise in interest rates does not produce a rise in savings then some components of the economy (firms or individuals) are not responding to the macroeconomic conditions or each other in a predictable way. Consequently we can only explain market failure by descending to lower levels of explanation, to microeconomics, possibly even to psychology and, since neuroeconomics is now part of economics, neuroscience. Much of microeconomics is now the psychology and neuroscience of non-rational choice for this reason.

This explains why, although macroeconomics does not reduce smoothly to microeconomics (once we abandon the idealising assumption of rational choice theory for individual cognitive processes), it is causally relevant. A change in savings rate is a change in the pattern of individual decisions and vice versa because individuals are microcomponents of the macroeconomy.

As with economies, minds, understood as cognitive economies, exhibit anomalies. A mind may respond to experience in an unusual way, for example by saying that the person who has that mind is inhabited by a lizard. When this type of anomaly persists we say the person is deluded. To explain this anomaly we need understand the interactions among the cognitive process which produce it and, ultimately, the influence on those cognitive processes of neural mechanisms. Procedural rationality need not be relevant to this type of explanation, unless the delusion resulted from failure of an explicit reasoning process at the personal level. More typically however it results from an upward cascade of interactions among a hierarchy of cognitive systems.

To explain conditions that present as irrational we do not necessarily to look for malfunction in a cognitive system that represents the norms of procedural rationality. In fact, insofar as procedural rationality allows us to understand persons it stands to the mind as macroeconomic theory stands to the economy.

How then should we proceed?

Theoretical Definition

I propose a theoretical definition of delusion that explains mutual manipulability between levels, including the personal level but which does not use normative notions proprietary to the personal level.

An introduction to the approach is provided by David Marr's definition of vision as the transformation of a monochrome two dimensional retinal array into a three dimensional representation of the visual scene. Strictly speaking, Marr did not intend this as a basis for interlevel neurocognitive theorising. He proposed it as the basis for a computational *model* which could in principle be implemented even in an artificial system (Marr, 1982). However research in human vision at all levels by disciplines ranging from computer science to molecular biology proceeded to use it to investigate relations of causal

relevance between visual experience, neural anatomy and activity and molecular level phenomena such as evolutionary changes to rhodopsin receptors. Such research transforms a cognitive model to a neurognitive theory. The point is not that the definition survived intact but that it provided a conceptual framework for simultaneous investigation of different levels of mechanistic functioning and their integration in a unified theory of vision. Even enactive theorists of vision who disagree with Marrians nonetheless debate with them about the causal relevance of mechanisms at different levels (Thompson et al., 2002).

Thus a good theoretical definition is one that ensures that investigators at all levels of a mechanisms function are *talking about the same thing*. It follows of course that such definitions are provisional and revisable. The current definition of delusion in the DSM IV in fact makes this integrative type of explanation impossible precisely because it deals in folk psychological notions which seem to apply exclusively at the personal level.

A false belief based on incorrect inference about external reality that is firmly sustained despite what almost everybody else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary.

It is not clear that this DSM definition is intended as a *theoretical* definition in my required sense rather than a clinical heuristic, or regimentation of everyday intuitions which builds in normative notion of rationality ("incorrect inference"). In the absence of any alternative, however, theorists often use the clinical definition as a basis for theoretical definition. For example in experiments that measure differences between delusional and non-delusional subjects on tasks that involve making inferences. These experimenters often conceptualise delusion as a failure to apply norms of inference or procedural rationality correctly and search for malfunctions in cognitive mechanisms that implement procedural rationality (Garety et al., 1991, Hemsley and Garety, 1986, Huq et al., 1988, Mujica-Parodi, 2000, Goel, 2004, Goel, 2007). These theorists use procedural rationality as a *measure of madness*.

One hope is that patterns of inference discovered in paradigms like these can be correlated with and perhaps ultimately explained by evidence about the neural and cognitive correlates of delusion. However before we commence the extensive and expensive search for non-existent entities we do need a way to ensure that all parties to the discussion are talking about the same phenomenon.

Even at the personal level almost all theorists have reservations about using procedural rationality as a measure of madness because it is not clear that the essential difference between non delusional and non-delusional people consists in their explicit or implicit grasp of procedural norms such as theorems of logic or probability or even informal heuristics which approximate them ("be consistent") (Bermúdez, 2000a, Stone and Young, 1997). Not only that but it has proved difficult to identity a cognitive and neural substrate for procedural reasoning. Instead of single systems which represent generative rules of logic or probability theory, procedural reasoning seems to involve the context-sensitive recruitment of other systems with different cognitive structures to fit a problem. Often "correct" or "incorrect" reasoning is an artefact of a cognitive heuristic activated by the context rather than the application of a procedural rule. For example if the problem has similar surface properties to a familiar problem type it will be solved by applying a schema (Goel, 2004, Goel, 2007, Goel, 2000).

Nonetheless, collecting evidence about the scope and severity of departures from ideal rationality in delusion has provided a large database of evidence and a number of useful hypotheses. In particular experiments within the procedural paradigm do suggest that the thought patterns of delusional subjects show entrenched *metacognitive biases*, even if ultimately the nature of those biases is not best conceptualised as failure of reasoning (Gerrans, 2001, Broome, 2004, Campbell, 2001b).

In any case the theoretical definition I propose provides a way to account for the origin and maintenance of those biases without invoking procedural norms. The aim is to identify cognitive processes at a level of description that identifies and explains the influence of neural processes on cognition and the way in which personal level phenomena emerge from those cognitive processes.

Delusions arise when

default cognitive processing unsupervised by decontextualized processing* is *monopolised* by *hypersalient information

Each of these theoretical terms, supervision, default cognitive processing , salience, and decontextualized cognitive processing describe systemic information-processing features of the human mind in a vocabulary which is "neutral" between personal level and lower level states. Default cognitive processing for example while it has a distinctive personal-level profile (by which it was initially detected) also has a distinctive neural-substrate (the default mode network (DMN) as it is known) which produces characteristic patterns of activation (Assaf et al., Broyd et al., 2009, Buckner et al., 2008, Fair et al., 2008, Gusnard et al., 2001, Horovitz et al., 2009, Kennedy and Courchesne, 2008, Mason et al., 2007, Pomarol-Clotet et al., 2008, Raichle and Snyder, 2007, Spreng et al., 2009, Whitfield-Gabrieli et al., 2009). The nature of the cognitive processes enabled by activity in these default circuits explains the distinctive profile of "default thought" at the personal level. For example different levels and patterns of activation in default circuitry, caused by low level neurobiological processes have predictable and explicable consequences at the personal level. The characteristic topics and patterns of thought of autistic and schizophrenic subjects are *predicted* by the differing baseline levels of activity in default circuitry, given the nature of the cognitive processing performed by that circuitry.

Similarly the concept of salience is neutral between neural cognitive and personal levels. Salient information, as the name implies, is that which is relevant. The mind has evolved mechanisms to ensure that expensive cognitive, behavioural and metabolic resources are allocated adaptively (Berridge and Robinson, 1998, Heinz and Schlagenhauf, 2010, Kapur, 2003, Kapur, 2004, Kapur et al., 2005, McClure et al., 2003). These allocations are reflected at the personal level concepts in like attention, concentration, working memory, cognitive control and executive function. These are folk-psychological expressions of the fact that cognition requires the allocation of cognitive resources to relevant information. Precisely which information is salient is a context-sensitive matter, which is why the salience system has evolved to support the flexible processing of information including the re-evaluation and recalibration of goals. Often even an extremely important train of thought or behaviour needs to be interrupted to deal with a new contingency. So one important role for the salience system is to help determine which information stays in the background so to speak at any time and which is allowed to monopolise cognitive resources.

Importantly, as with the DMN, the resource allocation role of the salience system can be detected and explained at different levels, as increased neural activation, as prioritisation of cognitive processes as well as in personal level as biasing of attention and executive function. It is only once the nature of the salience system is understood, mechanically and computationally that the role of dopamine in delusion and normal cognition becomes explicable as a mechanism which adaptively biases the integration of cognitive functioning across widely distributed circuits.

Delusions then reflect a spectacular misallocation of cognitive resources. Default thoughts (those produced by default cognitive processes) and associated patterns of thinking become extremely salient making it difficult for other information to attract processing resources. In particular so-called decontextualised thinking processes remain relatively inactive in delusion, which leaves the subject at the mercy of entrenched patterns of association in her default system.

This conceptualisation of delusion may not map directly to folk psychological conceptions of delusions as false beliefs caused by faulty inference. This is in fact a positive feature because, as we saw, this doxastic definition of delusion makes them opaque to multilevel investigation.

Another positive feature of this account is that it allows aspects of delusion, obscured by the incorporation into the clinical definition of the folk psychological conception, to come more clearly into focus. In particular the ambivalent epistemic attitude to experience of delusional subjects, which troubles doxastic theorists and occupied the early phenomenologists, can be explained in terms of the personal-level properties of default cognitive processing.

Although the account does not identity delusions as irrational beliefs it does help explain their irrationality measured against canonical standards of consistency. The default circuitry which produces delusions did not evolve to test beliefs for consistency with evidence and other beliefs. That kind of "reality testing" as the literature calls it requires the activation of circuits involved in decontextualized cognitive processing. However consistent activation of those circuits and their fluent interaction with default circuitry is compromised in delusional patients.

What then are default and decontextualised cognitive processes and how are their interactions affected by the salience system? This requires a description of the basic structure of the processing hierarchy and the way it is modulated by the salience system at different levels and time scales. The next chapter describes the basic structure and the salience system as a background to the account of default thinking in chapter 4.

THE PROCESSING HEIRARCHY AND THE SALIENCE SYSTEM

The processing hierarchy

The mind is organised as hierarchical system to which uses representations of the world and its own states to control behaviour. According recently influential Bayesian theories of the mind all levels of the cognitive hierarchy exploit the same principle: error correction (Clark, 2012, Jones and Love, 2011, Friston, 2003). Each cognitive system uses models of its domain to predict its future informational states given actions performed by the organism. When those predictions are satisfied the model is reinforced, when they are not the model is revised or updated and new predictions are generated to govern the process of error correction.

Discrepancy between actual and predicted information state is called "surprisal" represented in the form of an error signal. That signal is referred to a higher-level supervisory system, which has access to a larger data base of potential solutions, to generate an instruction whose execution will cancel the error and minimise "surprisal". If my serve goes into the net I must aim higher.

Surprisal arises when informational states are *not predicted* by the model. This can range from losing one's balance to having a marriage proposal rejected or failing to find weapons of mass destruction. In each case a prediction is not satisfied and the model that generates the prediction must be revised. (According to prescriptive theory. In practice,

as we know, evidence is often ignored, reinterpreted or discounted). In this sense a scientist explaining some discrepant evidence is doing the same thing as the oculomotor system controlling the trajectory of a limb: using and revising a model according to the degree of predictive accuracy it produces.

The essence of Bayesian conceptions of cognitive processes, understood as models, can actually be separated from the question of whether human cognitive systems implement Bayes' theorem. i.e. whether Bayesianism is a true theory of the way in which the mind detects and correct errors. Predictive coding theories of human cognition (Friston, 2003, Hohwy et al., 2008) treat Bayesianism as a neurocognitive theory.

Much of human cognition is quite automatic. The detection and correction of error occurs at low levels in the processing hierarchy at temporal thresholds and using coding formats that are opaque to introspection. Keeping one's balance and phoneme restoration are examples. We have no introspective access to the cognitive operations involved and are aware only of the outputs. This is the sense in which our mental life is *tacit*: automatic, hard to verbalise, experienced as fleeting sensations which vanish quickly in the flux of experience. This is the "Unbearable Automaticity of Being" (Bargh and Chartrand, 2003) .

However we are not complete automatons. At the higher levels of cognitive control surprisal is signalled in experience or explicit thought: formats available to metacognitive systems which evolved to enable humans to reflect and deliberate to control their behaviour. These metacognitive mechanisms released us from cognitive automatism: control by rigid routines automatically initiated by encounters with the environment (Lieberman et al., 2002, Proust, 2006, Proust, 2007, Miller and Cohen, 2001).

.

Delusions arise at the highest levels in the hierarchy when agents reflect on salient information referred by lower level systems. Thus to explain them we need to explain the basic computational architecture of referral and supervision. In particular we need to

explain how information becomes *salien*t, that is to say becomes the object of cognition at the imprecise borders between controlled and automatic cognition.

In fact the mind has evolved a salience system that manages the flow of information throughout the hierarchy, ensuring that cognitive resources are allocated to minimisation of the most important categories of surprisal. At the highest levels in the hierarchy we are aware of this resource allocation as cognitively-depleting directed attention and concentration . Resource allocation systems at these high levels of executive control recapitulate very basic mechanisms of salience which evolved to bias automatically controlled behaviour (Berridge and Robinson, 2003, McClure et al., 2003, Berridge and Robinson, 1998).

In order to explain how these mechanisms work we need a cognitive theory that displays the relevant properties of the salience system at all levels from automatic and reflexive to deliberate and reflective. That theory is provided in the form of a distinction between weight and activation based processing in neural networks which allows us to show how the neural mechanisms of salience influence cognition. In fact without such a computational/cognitive theory it is difficult to discern exactly what these systems are doing.

A Computational Framework.

It is hard to see what computational properties are common to automatic processes. At the level of representation and algorithm, as David Marr put it, what have language-parsing, vision and expert musicianship in common? The same apparent cognitive heterogeneity infects controlled processing. What do designing bathroom fixtures, mentally rehearsing answers to exam questions, thinking about one's holidays, plotting a revolution, solving algebra problems and being deluded have in common?

Neural processing provides the wrong level of description to allow us to distinguish automatic and controlled processing. At that level we see connectivity and conductivity

in neural circuits, not the representational architecture and cognitive processing of those circuits.

A standard distinction starts by defining automatic systems as modules: more or less autonomous subsystems which depend on specialised neural circuits to process input from the sensory periphery according to standardized cognitive routines(Bergeron, 2007, Coltheart, 1999, Fodor, 1983, Currie, 2000c). Controlled processing is a more "central" process integrating and manipulating information from the modules. Debates about the nature, structure, discreteness, flexibility and extent of modularity and central processes are continuing issues within cognitive science. Everyone agrees that there is cognitive and neural specialisation which enables different degrees of automaticity, and that there is also higher-level, more flexible, cognitive integration and metacognition, culminating in conscious explicit thoughts of the linguistic type conveyed on this page. Other than that all is controversy (Karmiloff-Smith, 1994, Plaut, 1995, Anderson, 2007).

For our purposes we can abstract from the controversy while retaining the importance of the distinction by recasting it as a distinction between weight-based and activation-based processing in neural networks (Arbib, 2003, Maia, 2005, McClelland et al., 1995, Munakata and McClelland, 2003, Plaut, 1995, Rumelhart, 1998, Stein and Ludik, 1998). A neural network is an abstract description of the properties of a neural circuit that treats it as a device that maps an input to an output. For example the visual system is a neural network (more accurately an interconnected set of neural networks) that takes input from the retina (primarily) and transforms it to a representation of a visual stimulus.

In a neural network the input is processed by a layer of input units. Input layers are like pixel arrays whose individual pixels are activated by electrical signals. Each unit in the input layer is connected to units in another middle or "hidden" layer so that activation in the input layer propagates to units in the hidden layer(s). Each unit in the hidden layer is "tuned" so that its activity is dependent on a threshold of activity propagated by units in input layers. The probability that a unit will become active given a level of input is the "weight" of a connection between units. Units in hidden or middle layers are connected

to units in an output layers in the same way. By adjusting the weights between units, different mappings from inputs to outputs can be produced. For example a network trained to recognise faces will represent facial features as activation patterns across input layers and identities as activation patterns across output layers. Learning consists in adjusting weights in hidden layers until the right mappings are produced (O'Reilly and Munakata, 2000).

A neural network is a simplified, idealised and abstracted *model* of the computations performed by actual neural circuits. A crucial aspect, which mimics neuronal properties, is that activity in individual units is modified only by interactions with units to which they are directly connected. Any global changes to the system have to be achieved by modifying local informational properties. The art of the modeller is in producing an algorithm for setting weights between units which mimics the transformations from input to output of an actual cognitive process. Algorithms can mimic feedback, recurrent and serial processing between circuits and learning and evolutionary processes. Of course as we noted earlier, transforming a model which reproduces input output mappings of a cognitive system, to a theory of the way that system works requires a description of the neurocomputational mechanisms which implement that model (Arbib, 2003). The art of the neurocognitive theorist is finding an algorithm which describes the way actual neural circuits stabilise their cognitive properties in response to constant modulation by the other circuits to which they are connected (O'Reilly and Munakata, 2000, O'Brien and Opie, 1999).

There are two features of neural networks which make them useful for explaining the effects of the salience system on controlled processes.

The first is the distinction between activation in a layer and the weights of the layer (O'Reilly and Munakata, 2000). This corresponds to the properties of a neural circuit when innervated or at rest, which is the same as the distinction between occurrent or dispositional properties. The face recognition system has the dispositional property of producing judgements of identity when shown a face. In a neural network model this is

achieved by changing the weights in hidden layers (either by hand or via an evolutionary or learning algorithm) so that facial features and configurations presented as inputs are mapped to an identity judgement (Young and Burton, 1999). However the system requires an input in order to start the mapping process. It needs an input representation of a face to map to the output of identity or familiarity.

It is in this sense that weight-based processes are best described as *reflexive.* When presented with an input they automatically initiate cognitive processing mapping it to an output. For example when presented with a face the face processing system cannot help identifying it and parsing it for emotional expression. This is what Fodor (1987) called the mandatory aspect of modular processing.

The second point is that the weights in hidden layers function as the system's memory. The face recognition system does not store a separate representation for Madonna, Nelson Mandela and Julius Caesar. Rather weights are set so that input representations of the faces of Madonna, Julius and Nelson are mapped to the right outputs. Thus the weight structure of the hidden layers represents *all possible mappings* from input to output. Which mapping is produced depends on the activity in the input layer and the weights of the hidden layers. Input and output representations are discrete, the dispositional memory of the system is holistic.

It follows that the domain of a network (which inputs it can process) and its degree of flexibility in processing depends on the weight structure. If the weights are fixed and cannot be changed the input-output mappings cannot be changed. This is why it makes sense to think of the circuits that perform reflexive processing as neural networks with inflexible weights in their hidden layers. This inflexibility is an advantage when processing stereotypical stimuli (a visual system which produced different representations of the same stimulus as output would not be fitness -enhancing). Evolution and learning are thus mechanisms for setting weights at different time scales. It also smooths output by abstracting away from irrelevant variation. Colour constancy for example is achieved by compensating for variation at the input level. Similarly we

continue to recognise faces under very different conditions because the weights in hidden layers filter out irrelevant variation to produce an invariant " structural description" as it is known.

The inflexibility of weights in automatic systems also models the way non standard inputs cannot be mapped smoothly to outputs: if the weights cannot "recognise" the input, activation in hidden layers will not settle quickly and will not propagate to a standard pattern in output layers. The result is an error signal or non-standard activation pattern in output layers accompanied by unusually sustained activity in hidden layers. This is the "surprisal" which is the raw material of Bayesian cognitive processing.

These error signals propagate to error higher-level systems which take them as input and the process iterates. At the highest levels those signals are sensed as experiences which can become the focus of attention and initiate reflective deliberative responses.

Thus, computationally, an automaton is an integrated collection of weight-based processing modules. Outputs of one system (or a part of one system) such as visual perception serve as inputs to another (or part of another) such as face recognition. Furthermore *the system depends on the presence of stimuli in the environment to initiate weight- based processing*. When there are no stimuli there is no activity in input layers and activation in hidden and output layers decays. These properties of weight-based systems reflect important feature of posterior neural circuits specialised for automatically managing encounters with the environment (Lieberman et al., 2002, Lieberman et al., 2004). The fact that there is no real data/process distinction, no discrete representation and the speed of the input/output mapping account for the tacit nature of automatic processing. When I recognise a familiar face my face recognition system has mapped the input from the visual system to identity, and other automatic response systems, within a few hundred milliseconds. Stages of the mapping process serve as inputs to other systems such as recognising emotions or attributing intentions. There is recurrence and feedback between these networks but it all takes place at speeds and using holistic coding formats that make automatic processes opaque to introspection.

Of course humans are not entirely automations. Precariously superimposed over our automatic systems are a variety of supervisory systems that confer capacities for higher level control. Controlled processes are best modelled as higher-level "supervisory" neural networks which can maintain and manipulate activation in output layers of weight-based systems in the absence of activation in input layers. When I see the face of the Regius Professor of Gender Studies retinal input is automatically mapped to output layers of the visual system producing a visual image. If I imagine seeing him that visual image of his face is produced in the absence of the professor himself. This requires the " top down" production of the relevant pattern of activation in output layers of the visual system. If I want to mentally compare the appearance of the Regius Professor before and after gender reassignment surgery I need to reconstruct and compare two images. If I want to compare his CV to that of other candidates for the vice chancellorship I need to construct a different, more abstract type of representation. These are not automatic but voluntary, not reflexive but reflective, tasks requiring the integration and evaluation of information not currently perceptible and therefore cannot be accomplished by weight-based processing.

This is the type of processing the prefrontal cortex evolved to manage by maintaining and manipulating outputs of posterior systems *in the absence of the stimulus* (Fuster, 2008, Gilbert et al., 2005, Knight, 1999, Miller and Cohen, 2001, Waltz, 1999)

Thus an essential difference between pure automatic and controlled processes is the degree of stimulus dependence. In controlled processing the outputs of different weight-based systems are reconstructed by memory or imagination, simultaneously held in working memory, metarepresented, compared and manipulated into novel combinations. Hence controlled processing is *activation-based* processing because it selectively maintains activation in neural circuits which would otherwise decay as automatic cognition continues its Sisyphean struggle to manoeuvre the organism through the world.

The salience system and reward prediction

Salience is the functional concept that describes preferential allocation of resources to the processing of *relevant* information. With the contrast between activation and weight-based processing in place we can begin to explain the mechanisms which make information salient to controlled processing.

The explanation of salience is an instance of the integrative approach to mental phenomena. It combines the insights of formal learning theory, are expressed mathematically, with computational models of learning in neural networks to construct models which solve a crucial cognitive problem: the allocation of resources to relevant information. However the *models* which solve that problem remain models: abstract descriptions of an information processing architecture. In order to transform those models to theories of actual human (or animal) cognition we need to describe the mechanisms which implement them. In that process of description the theory itself is revised in the face of the evidence. The human mind was not designed *a priori* by a mathematician but evolved, and the operations of its salience system reflect that history.

This process of theoretical development is a success story for the multivel integrative approach. Predictions about the personal level effects of intervention at molecular level (e.g. the blocking of dopamine receptors or the introduction of dopamine agonists at specific sites) can be modelled then tested. And in fact the explanation of the salience system proceeds at all levels. Personal level phenomena such as addiction and delusion characterised in personal level terms like “desire” and “faulty inference” are explained in terms of the biasing of the flow of information in the control hierarchy.

Because all organisms, even the simplest unicellular automatons, need to allocate energy efficiently rudimentary salience systems abound in nature. In the mind of an organism like a squirrel the salience system ensures that its mind is captured by salient stimuli such as nuts and foxes and remains focused on them until an adaptive behavioural response is produced and learnt. The phenomenon is baptised *incentive salience* since the salience system is operating to make perceived objects motivational magnets. A crucial point about motivational magnetism is that the adaptive response itself is not immediately

rewarding. Nut consumption and fox evasion are intrinsically rewarding, running and climbing are not. The role of an incentive salience system is to install these instrumental behaviours (Berridge and Robinson, 2003). Thus, for example, were we to plant electrodes in the brain of a foraging squirrel we might initially see activity in salience systems associated with eating nuts (or dodging foxes). Over time this activity would be replaced by activity in the salience system associated with initiation of successful foraging or evasion routines. The role of a salience system is not to reward success, but to *predict reward* for an organism (Schultz, 1997, Berridge and Robinson, 1998, Gottfried et al., 2003, Heinz and Schlagenhauf, 2010, McClure et al., 2003, Egelman, 1998, Smith, 2006).

While much, perhaps more than we like to think, of human life is squirrel-like, humans are not entirely stimulus-response automatons hostage to ancient incentives (although a look a the Mens and Womens shelves of the newsagent is grounds for pessimism). The human capacity for agency depends on the evolution of elaborate neocortical structures which enable humans to inhibit automatic responses, select alternative responses and, ultimately, to construct and manipulate simulations of alternative scenarios in order to plan longer term responses and to use these representations to control behavior. Agents have evolved dedicated prefrontal neural structures to enable the construction and maintenance of transient distributed circuits linking the prefrontal and posterior areas that implement these simulations. These metabolically expensive and fragile circuits represent the information needed for stimulus-independent controlled cognition.

Controlled cognition uses the same salience system as automatic cognition to allocate resources to relevant information. The mind has has co-opted and adapted existing mechanisms at each stage in the evolution of greater cognitive control. The essential component of the salience system is the dopamine (DA) system which projects from origins in the Ventral Tegmentum Area (VTA) and basal ganglia (BG) throughout prefrontal, posterior and limbic areas. Neural network models which preserve biochemical parameters of the VTA prefrontal cortical (PFC) and BG-PFC networks involved in activation-based processing have been be used to test the effects of

dopaminergic innervation on controlled processing (Braver et al., 1999, Braver and Cohen, 1999, Abi-Dargham et al., 2002, Durstewitz and Seamans, 2002, Egelman, 1998, Goldman-Rakic, 1997, Grace, 1991). These experiments confirm its role as a salience system and provide a deeper theoretical basis for understanding the correlations between dopamine dysregulation and the phenomenology of delusion.

In these models dopamine enhances the signal to noise ratio (SNR) between communicating neural circuits. It does so via the interaction of at least two types of DA action. Phasic DA, delivered in short bursts, binds to D2 receptors on the postsynaptic membrane. It is rapidly removed by reuptake from the synaptic cleft and acts quickly. It is described as producing gating effects: determining which representations are allowed to interrupt and enter controlled processing. Of course "gating" is a spatial metaphor: entry into controlled processing refers to levels of activation sufficient to capture and retain attention, monopolise working memory and executive functions. A pattern of activation amplified and reinforced by phasic dopamine activity out-competes other patterns of activation.

Tonic DA, which acts over longer time scales, accumulates in the synaptic cleft and binds to presynaptic DI autoreceptors triggering reuptake. This contrast between tonic and phasic activity is a ubiquitous neuroregulatory strategy. Tonic levels of a neurochemical are delivered and maintained at steady levels by slow regular pulses of activity. Phasic activity is intense staccato and short-lived, interrupting the ongoing activity maintained by tonic levels.

Phasic and tonic DA are thus antagonists and have different effects on the circuits they afferent. Phasic DA, acting on PFC-posterior circuits produces a gating effect. It allows new activation patterns in the PFC-posterior circuitry to be formed, allowing representations of new stimuli. Tonic DA maintains an occurrent activation pattern allowing a process to be sustained against interference or competition. The hypothesis follows, and is consistent with neural network models, that the balance of tonic and

phasic DA is responsible for the rate of turnover of representations in the PFC-posterior networks (Grace 1991).

The effects are captured intuitively in the folk psychological concept of concentration. Too little phasic activity leads to perseveration: of behaviour in automatons, of patterns of thought in agents. Too much is usually evidenced in impulsivity in automatons and distractability in agents.

Controlled cognition needs to strike the right balance between perseveration and distractability according to the value for the organism of maintaining a train of thought or switching to deal with new information. Dopamine delivery is the basic mechanism which maintains this balance. Together phasic and tonic dopamine provide a mechanism for the updating and maintenance of representations in working memory by selectively activating target neural assemblies (Arnsten, 1998).

> Tonic DA effects may increase the stability of maintained representations through an increase in the SNR of background versus evoked activity patterns. In contrast, phasic DA effects may serve as a gating signal indicating when new inputs should be encoded and maintained (Braver 1999, p317).

Computationally we can say that gating effects influence activation patterns and hence determine which representations are available for controlled processing. Learning effects change the weights in neural networks, the equivalent of long term potentiation required for long term memory. The actual mechanism in neural circuits is Hebbian modification of synaptic connections: when a pattern of activation is repeatedly sustained the relevant synaptic pathways are strengthened via a cascade of chemical changes which ultimately change the receptive properties of target neurons. These learning effects are produced by the initiation of mechanisms of long-term potentiation by the presence of dopamine at receptors on the cell surface.

To summarise: controlled cognition involves flexible construction and maintenance of representations whose neural bases are transient patterns of activation across adistributed PFC-posterior circuits. Which representations are evoked, maintained and ultimately encoded in long-term memory depends on dopamine-modulated salience (Goldman-Rakic, 1997) .

Salience and the adaptive critic

Computational modelling captures this role for the dopamine system in what are called Adaptive Critic architectures (Barto, 1994, Singh et al., 2005). The Adaptive Critic is a cognitive system which learns through feedback which contingencies are rewarding in order to order modulate the activity of another system (the Actor) which needs to allocate processing resources adaptively. For example a squirrel choosing between oak and pine trees needs to develop a preference for the one which has the food at the top. Climbing the tree is unrewarding instrumental behaviour which needs to be installed in order to produce the ultimate reward. The mechanism is dopaminergic enhancement of activation in circuits that link instrumental behaviour (tree climbing) to the perception of the tree.

Thus computational neuroscience describes the dopamine system as a reward *prediction* system. Its role in automatons is to install potentially rewarding behaviour by conferring incentive salience on perceived objects. (Schultz et al 1997; Berridge and Robinson 1998; Braver 1999; Berridge and Robinson 2003; Fiorillo and Schultz, 2005; Kapur 2003; Clure et. al. 2003; Smith et. al. 2006). 2004).

Computationally we would model this as setting weights in a circuit which linked perception of oak trees to the relevant exploratory behaviour. The adaptive critic is the name of the component of a model that sets these weights by ensuring that perception of oaks initiated foraging.

The crucial point about Adaptive Critic architectures is that the reward is in the future Adaptive critics in effect teach a system to produce the intrinsically unrewarding instrumental activity for the sake of a distant reward.

Rats and squirrels are (largely) automatons but the value of the adaptive critic model is not restricted to explaining the behavioural preferences of rodents. In fact controlled cognition is so metabolically expensive and potentially computationally unconstrained (so many potential problems, so many potential solutions) that it requires a salience system. For example whether or not to maintain an item in working memory or continue to pursue a train of thought depends on the potential value of allocating resources to that representation. In the exam it is not a good idea to fantasise about lying on the beach in Tahiti, but at the travel agent it may be the most relevant use of metacognitive resources. Imagining dying of skin cancer is hypochondriacal time wasting in the office but may be a good idea on the beach.

Rats in mazes and humans in life (sometimes metaphorically described as a “rat race”) face the same problem: to allocate resources adaptively. This requirement and structure of the solution have been demonstrated mathematically and modelled computationally. However a formal solution to a cognitive problem is not a demonstration that the brain implements that solution. It remains a model of a cognitive process. To transform that model into a cognitive theory requires a description of the interactions of the dopamine system with the circuitry it innervates in a variety of experimental and clinical conditions including delusions.

The reward prediction framework predicts that the balance of tonic and phasic dopamine delivery would modulate the salience of representations at different levels, influencing learning, memory, planning and decision and motivation. Furthermore since *unpredicted* activity, which constituting surprisal, is most salient and likely to be referred to controlled processing phasic dopamine activity which interrupts ongoing activity should be associated with novelty.

These predictions are borne out in single neuron studies of the VTA of rats in a variety of paradigms. For example in a conditioning paradigm in the learning phase dopamine neurones fire for the reward (Waelti et al., 2001, Montague et al., 1996, Schultz, 1997). As the association is learned, firing for the reward is reduced and dopamine neurones fire for the instrumental behaviour. In other words they "predict" reward. (Waelti et. al. 2001 p43).

Not only that but dopamine neurones fire when predicted rewards do not arrive " in a manner compatible with the coding of prediction errors." (Waelti et. al. 2001 p43) and also respond to novel attention-generating and motivational stimuli.

In other word it seem that the role of the dopamine system is to focus cognition on relevant stimuli. Events which are consistent with models produce less phasic activity in the dopamine system, than novel (i.e. unpredicted) and affectively valenced (good or bad for the organism) events. Not only that but once the associations are learned dopamine functions as a reward prediction system, increasing firing for instrumental activity but reducing firing if the reward does not arrive.

Current understanding of the reward system suggests that higher-level control systems preserve the basic structure of lower level reward prediction. For example neurones in the ventromedial prefrontal cortex, a crucial way station between automatic and controlled cognition and hub of the default system are innervated by the dopamine system. They exhibit very similar properties to dopamine neurones in the Ventral Tegmentum Area classically identified as a locus of reward prediction. Namely they exhibit the same temporal properties as dopamine neurones for affectively valenced stimuli.

In effect the ventromedial prefrontal cortex recapitulates the structure of automatic reward prediction *in the absence of a stimulus* (Gottfried et al., 2003). When we lie on the beach we experience pleasant sensations. When we are in the travel agent our ventromedial prefrontal cortex predicts for us that a trip to Tahiti will be rewarding by associating a rewarding experience with an imaginary episode of beach-lie-ing.

The fact that controlled processing recapitulates the structure of reward prediction and exploits the same ancient mechanisms is unsurprising. Stimulus independent thought is in effect cognitive foraging: a search through representational space for rewarding intentional objects and should obey the same rules as foraging for physical objects in physical space. As new cortical systems were layered over older ones enabling higher levels of control and more abstract forms of representation they inherited the same problems of resource allocation and the same mechanistic solutions.

Dopamine and Delusion

Against this background the explanation of the role of the dopamine system in delusion turns out to be not so much a discrete psychological puzzle but a piece in the larger puzzle of understanding the relationships between lower and higher level control systems and the salience systems which modulate them.

A correlation between dopaminergic activity and psychosis has been well established for many years but transforming correlation to explanation required the integration of many levels of description of its effects via the idea that it functions as and essential part of a salience system.

The “dopamine hypothesis” of psychosis starts from the facts firstly, that antipsychotic drugs are dopamine antagonists, whose effects are achieved by blocking the action of dopamine. The second fact that is dopamine has a role in triggering or mimicking psychosis, evidenced by hallucinogenic drugs and heightened dopamine synthesis during psychosis (Crow, 1980; Jones and Pilowsky,2002; Heinz and Shlagenhauf 2010). Furthermore abnormalities of the dopamine system particularly the reception of phasic dopamine have been demonstrated in psychotic disorders such as schizophrenia.

Despite these correlations the “dopamine hypothesis” initially stalled because of the lack of a direct relationship between dopamine levels and psychosis.

However the computational interpretation of the role of dopamine suggests that what matters is not the absolute level of dopamine. The *balance* of tonic and phasic dopamine and its context sensitive delivery rather than absolute levels, together with the distribution and activity of receptors in distributed prefrontal-posterior circuits make the difference. For example Ritalin (methylphenidate) increases dopamine levels and thus might be thought to be the last thing impulsive and highly distractable children need. These children are already hypersensitive to novelty. However Ritalin increases *tonic* levels of dopamine in the prefrontal cortex, making the cognitive activities of these children less vulnerable to interference by bursts of phasic dopamine which signal a salient new stimulus. For children with neurotypical levels of tonic dopamine, distribution and sensitivity of D2 receptors, the same delivery of phasic dopamine plays its proper gating role in redirecting attention to salient stimuli and organising executive functions adaptively. In other words these children strike the right balance between perseveration and distractability partly because they have the right balance between tonic and phasic dopamine delivery. Their salience system is correctly calibrated.

Recent work on the dopamine hypothesis of schizophrenia has concentrated on its role in the salience system, comparing levels of phasic dopamine delivery in conditioning and learning paradigms to that of normal subjects. The basic idea is that in psychosis is that the "wrong" representations become salient and relevant novel information is not processed. At low levels this is reflected in attentional deficits, at higher levels in failure to allocate metacognitive function appropriately. At all levels what counts as "surprisal" (prediction error) is different for the schizophrenic mind as a consequence of abnormalities in the way the salience system works. Summarising a range of studies Heinz and Schlagenhauf (2010) express a developing consensus:

> the blunted difference between relevant and irrelevant stimuli and outcomes may reflect chaotic attribution of salience to otherwise irrelevant cues, an interpretation that is in accordance with the idea that chaotic or stress-induced dopamine firing can interfere with salience attribution in schizophrenia. (p477)

The salience interpretation of the of the dopamine system theory provides a unifying explanation features of schizophrenia, including the characteristic phenomenology of the prodromal period in which subjects feel that events or objects are extremely significant and/or that their senses are hypersensitive. As Shitij Kapur points out transient episodes of this nature are not abnormal but in delusional subjects dopamine dysregulation ensures that their hypersalience gives representations of objects or scenes a halo of significance and ensures that they dominate attention working memory and executive function (Di Forti et al., 2007, Moore and Grace, 1999, Abi-Dargham et al., 2002, Grace, 1991, Howes and Kapur, 2009, Braver et al., 1999, Broome et al., 2005). As Heinz and Schalgenhauf put it:

> dopamine dysfunction may be particularly prominent during the early stages of schizophrenia before delusional mood is transformed into fixed and rigid patterns of delusional explanatory models; the model implicitly rests on the assumption that dopamine firing can be increased by environmental stress (p 474)

For Kapur psychosis originates at the border between automatic and controlled processing when information becomes hypersalient, attracting attention and commanding further prefrontal resources. He describes dopamine as " the wind of psychotic fire" ensuring that activity in circuits referring and processing delusion-related information increases to levels which make reallocation of resources to non delusional information impossible for the psychotic subject.

The idea that the mind is an hierarchical control system whose information processing depends crucially on a salience systems has an important implication for the explanation of delusion. Those experiences which signal surprisal will naturally dominate high-level cognition since it evolved to enable adaptive responses to problems that exceed the processing capacities of lower level systems. The salience system will interact with the neural circuits that refer these problems, amplifying their activity (increasing the "gain" is the technical term in neural network theory) and thereby making them the focus of attention. Another important feature of this explanation is that it suggests that delusions

are far more likely to arise for surprisal which cannot be cancelled from the top down, so to speak. Automatic perceptual and affective systems which continue to generate and obstinately refer error signals will produce experiences which monopolise higher level control systems (Corlett et al., 2009a, Corlett et al., 2010, Corlett et al., 2009b, Pankow et al., 2012). Delusions are a particular form of this monopoly. However this also suggests that delusions may share some features with other conditions in which higher-level cognitive control systems are hijacked by dysfunctional mechanisms of referral and supervision. Obessional or compulsive patterns of thinking are examples as are entrenched patterns of thought involved in mood disorder.

Within this framework trying to distinguish whether the thoughts of someone with extreme depression or obsession are delusional in the same way as someone with paranoid delusions may not be productive. A better approach is to trace the pattern of referral and supervision that generates the salient experiences and thoughts in each particular case.

In any case it will never be sufficient to show how surprisal, in the form of salient experience is generated. We also need to know why it is not cancelled and is often incorporated into a delusional train of thought. A delusion is a *response* to that surprisal "a top down cognitive phenomenon that the individual imposes on these experiences of aberrant salience in *order to make sense of them*" (Kapur, 2003, p15). Once adopted, the delusion "serves as a cognitive scheme for further thoughts and actions. It drives the patients to find further confirmatory evidence – in the glances of strangers, the headlines of newspapers and the tiepins of newsreaders" (Kapur, 2003, p16). The effect of hypersalience is to entrench the delusion.

Applications

The description in this chapter of the processing hierarchy and the salience system has been quite general. It applies to the mind in any mode. However it is important to the discussion of specific delusions (of misidentification and of alien control) in the

following chapters because those delusions are the responses of a higher-level control system, the default mode network, to signals of surprisal constantly referred from automatic systems. In delusions of misidentification the surprisal is produced by a mismatch between facial appearance, affective response and identity. In delusions of alien control the surprisal takes the form of a mismatch is between predicted bodily movement and an actual movement. In both cases the default systems generates a delusional response that makes the obstinate referral of surprisal intelligible from a subjective perspective.

An interesting feature of these delusions, and delusions in general, is that the delusion itself is not referred to higher levels of cognitive control in order to resolve the conflict between the delusion and other information such as background beliefs or the testimony of others. Or in some cases it may be referred but it cannot be cancelled by a command backpropagated from systems which rationally evaluate beliefs for their coherence and consistency.

There is evidence that the normal relatively smooth interface between the default system and higher-level supervision is disrupted in delusion. In schizophrenic delusion for example the salience system seems to be in overdrive, targeting the default systems, allocating all cognitive resources to the default thought and starving higher levels of control of the resources they need to deal with the delusion. In other delusions higher level systems are lesioned or compromised.

The precise mechanisms in each case need to be described in more detail and this chapter provides the necessary context. However before we discus particular delusions we need to examine the nature of the default mode network and its role in the control hierarchy.

THE DEFAULT MODE NETWORK

Simulations as narrative elements

The main task of this chapter is to explain the properties of the "default mode network" (DMN) the system that is active and relatively unsupervised in delusion. The default mode of the DMN is a resting state of a powerful simulation system that evolved to allow humans to simulate experiences in the absence of an eliciting stimulus. This capacity released us from the stimulus bound present, allowing us to recall past experiences and imagine possible futures. Mental time travel, as this type of simulation process is now called is an adaptation for planning. We can review and preview possible scenarios and experience characteristic emotions before committing to action.

These scenarios have a narrative structure that can range from very minimal to richly structured and elaborate. In mental time travel narrative elements are highly organised by an overarching goal. Imagine mentally rehearsing a journey or a speech for example. When not organised for problem solving the DMN reverts to the screensaver mode we experience as daydreaming or mindwandering. In this mode there is no overarching goal to provide narrative structure.

The DMN gets its name from the fact that its was detected and mapped in the resting state: daydreaming is what the mind does in default mode in between episodes of simulation generated for cognitively demanding problem solving. However focusing on the default mode in isolation turns attention away from three of its features that are essential to understanding its role in delusion

(i) it is a simulation system

(ii) the simulations are fragments of autobiographical/personal narrative

(iii) it is supervised by, and anticorrelated with, circuits specialised for decontextualised processing

Although the simulatory capacities of the default network can be understood in isolation it is best understood in terms of its role in the processing hierarchy: supervising lower level, more automatic, processes and supervised by higher level, decontextualized, processes. We explain its properties in different conditions. The first is alert waking cognition in which it is in communication with other perceptual and executive systems. This is familiar condition in which for example we rehearse possible future and actual past experience to help decide where to have lunch, send children to school or who to marry.

Another condition that displays the properties of the default network is dreaming. In dreaming the default network is disconnected from both the sensory periphery of the mind and from high-level control systems deactivated during REM sleep. As a result dreams are (relatively) cognitively incoherent. They are "dramatic simulations of our conceptions, concerns, and interests that occur when a specific constellation of neural regions is activated *in a context where there is no engagement with the external world*." (Domhoff, 2011)

Once these conditions are described we turn to the way in which the default network functions, relatively unsupervised, in delusion. Delusions are not dreams, or even waking dreams, but they have *some* of the properties of dreams because they depend on activity in default circuitry. Equally they have some of the properties of "normal" waking

cognition. They are simulations triggered by sensory and perceptual inputs that are subsequently incorporated into the agent's psychology.

The importance of this idea is that the falsity, fixity and fabulous aspects of delusion result from the nature of default thought itself. Delusions are not *caused* by failures of reasoning, since default thinking is not intrinsically a reasoning process. Intrinsically it is an imaginative/simulative process precariously disciplined by the demands of narrative coherence. Neither imagination or narrative are essentially governed by norms of rationality. In this respect theorists who remark on similarities between dreams and delusions are correct. In neither case are the simulations tried before the tribunal of reality.

Mental time travel and the default network

As Daniel Gilbert once put it.

> We are the only animals that can peer deeply into our futures—the only animal that can *travel mentally through time*, preview a variety of futures, and choose the one that will bring us the greatest pleasure and/or the least pain. This is a remarkable adaptation—which, incidentally, is directly tied to the evolution of the frontal lobe—because it means that we can learn from mistakes before we make them. (Gilbert, 2004)

Gilbert here invokes the notion of *mental time travel* to capture the idea that imaginative rehearsal, using information gathered in previous experience is the essence of planning. Gilbert's work participates in a reconceptualisation of the nature of episodic memory (for previous experiences) and imagination as aspects of a unified capacity for simulation of experience. An important aspect of this reconceptualization is the idea that that episodic memory and episodic future thought, *prospection* as it is known, are processes which depend on the same neural systems. This view is supported by evidence that damage to neural circuitry known to be essential for episodic memory also impairs imagination.

What this means is that in the case of episodic memory we simulate a past experience (more or less accurately) and in imagination we simulate a possible experience event cueing the same machinery in different ways (Botzung et al., 2008, Maguire and Hassabis, 2011, Maguire et al., 2010, Spreng et al., 2009, Hassabis et al., 2007a, Hassabis et al., 2007b, Hassabis and Maguire, 2007a). We re-experience the past to pre-experience the future. As Schacter et. al. put it: "the medial temporal lobe system which has long been considered to be crucial for remembering the past might actually gain adaptive value through its ability to provide details that serve as building blocks of *future event simulations*" (Schacter et al., 2007).

These considerations help explain one controversy about the nature of default activation in REM dreaming. Given that dreams include episodes of personal history some have explained them in terms of the endogenous activation of an episodic memory system. However against this interpretation theorists like Solms, Schwarz and Maquet point out that dreams do not really present as memories (they have none of the phenomenology of being "re-experienced") but rather as streams of experience which juxtapose elements of previous episodes of the subject's life with episodes which, while infused with a feeling of subjective importance, have no connection with the subject's actual life history (Domhoff, 2005, Solms, 2007, Solms, 2000).

If however we reconceive of episodic memory as an aspect of a unified capacity for simulation via the endogenous activation of outputs of automatic perceptual and sensory systems the controversy takes on a different shape. Whether a simulated episode constitutes a memory cannot be determined by noting that it depends on "memory" circuitry since those same circuits participate in planning and imagination as part of the default network. Not only that but those same simulatory mechanisms can be endogenously activated in dreams.

Under executive control these simulative abilities provide a form of cognitive control unavailable to creatures whose cognitive repertoire is limited to automatic responses to

environmental stimuli. We do not need to actually perform an action to evaluate its consequences. We can simulate it in memory or prospection before proceeding to action.

Mental time travel is highly developed in, if not unique to, humans (Suddendorf, 2006) and monopolises our reflective waking life. This monopoly is maintained *even when we are not planning for a specific goal.* Rather than go to sleep between intervals of cognitively demanding and metabolically expensive mental time travel the human mind reverts to a *default state* while waiting for another salient problem to focus the simulation process. Rather than logout and reboot the mental time travel network we simply switch to screensaver mode. As long ago as the 1970s David Ingvar noted that frontal regions showed greater blood flow in resting states compared to states of perceptual attention and hypothesised that this indicated that the brain was not actually idle but engaged in spontaneous activity which involved prefrontal areas. A 1995 study by Andreasen drew a connection between autobiographical memory and resting states hypothesised to be a " mixture of freely wandering past recollections, future plans and other personal thoughts and experiences ". In another important study Shulman et. al. mapped cortical areas implicated in this default state by meta-analysis of PET studies involving active-passive contrasts and showed that a widely distributed cortical network was implicated in the rest condition. In 2001 this network was baptised the default network in three important papers which verified the role of medial prefrontal regions associated with "self referential processing" (Gusnard et al., 2001, Sajonz et al., 2010) (Whitfield-Gabrieli et al., 2011)in conditions where the mind is not engaged in goal directed activity (These studies are discussed in Buckner et. al. 2008)

Recent studies have complicated this picture slightly, arguing that the default network has internal structure which reflects the different roles it performs in active, mental time travel, mode. This idea picks up an earlier suggestion that it can function in "sentinel" mode, prepared to deal with incoming perceptual information or in "social simulation" mode, rehearsing potential social interactions. In both cases it is functioning as a screensaver, turning over representations in a low energy state, poised to resume goal-directed processing. As Broyd et. al. put it: the properties of the default network reflect

"low frequency toggling between a task-independent self-referential and introspective state [daydreaming] and an exteroceptive state that ensures the individual is alert and attentive to unexpected or novel environmental events." (Broyd et al., 2009).

Delusions as a "mixed mode" of the default network

In either case the default network is poised to respond to surprisal by creating simulations. In the absence of a focus for simulation it reverts to one of its default states. It follows that the "purest" or deepest form of default processing is sleep. In REM sleep we see activity in the default network combined with a high degree of cognitive disorganisation in the simulation process, reflecting the fact that exteroceptive channels and higher -level control systems which represent longer term goals and keep the default system "on track" are deactivated. Dreaming is a state of representational incoherence because the default network is neither dealing with a current problem nor poised to do so. In dreaming the default network is at the mercy of innervation by subcortical systems that imbue endogenously produced representations with salience and affective tone. The eeriness and significance of the dreamscape are a product not only of the incongruities the dreamer encounters but the fact that dream experience is felt to be salient and emotionally significant.

The claim that delusions are default thoughts is *not* the claim that delusions represent the mind in waking default mode: that is subjective daydreaming when not focused on a salient environmental or social goal. Rather the claim is the delusions arise when the default network performs the task for which it evolved. Namely simulating scenarios in response to salient information. So delusions arise when the default network is in active mode: receiving information from the sensory periphery and interacting with memory and other high level cognitive processes such as attention and executive function. Those interactions however have a different character in delusion because of a bias towards default processing of delusion-related information in the delusional mind.

There are some crucial differences between default processing in delusional and non-delusional minds. The first is the nature of the information that triggers activity in the default network. For example delusions of misidentification are default responses to abnormalities of processing in the face recognition system. In this case the delusion represents a default response to a quite specific experience whose content is reflected in the delusion. The second is hyperactivity and hyperconnectivity in that network. This *automates* the generation of the delusional thought pattern consequent on an initiating experience and reinforces that very delusional pattern. Crudely, once the subject has had the default thought "my wife has been replaced by an imposter" in response to the experience of misidentification, this default thought and its associations are triggered on the next occurrence of the precipitating experience. This type of reinforcement is very likely related to the third difference: the interaction of the *salience system* with default processing circuitry. The salience system allocates resources to the default processing of delusion-related information. The final difference, often related to the third, is the absence or corruption of *decontextualised supervision*. Because the delusion remains highly salient it is difficult to refer to higher-level processes for decontextualized evaluation. Alternatively decontextualised systems may be damaged or hypoactive which has the same effect.

There are different ways in which fluent interaction between default and decontextualised processing can be impaired, reflected in different profiles of delusional disorders. For example in some of the neuropsychological delusions (of misidentification) discussed below and in subsequent chapters the circuits responsible for decontextualised supervision are lesioned. Consequently the patient is at the mercy of a default thought triggered by an error-signal in a lower level system. Interestingly in these delusions activity in the default network may not be affected by hyperactivity in the salience systems, which may be why, by the standards of delusions, they are cognitively coherent, monothematic and circumscribed. In other cases the default network may be hyperactivated via its interactions with the salience system which makes it both hyperassociative and unable to be deactivated to allow for decontextualised supervision.

This seems to be the case in full-blown psychoses and to a lesser degree in polythematic and elaborate delusions.

Thus delusion while they reflect the intrinsic properties of the default network, are not entirely explained by those properties. They also depend on the way the default network is cognitively embedded. For the moment, however, we concentrate on those properties. First in waking cognition, and then in dreams.

The first person perspective and decontextualisation

Default circuitry evolved to enable humans to make salient information intelligible from the perspective of the subject. Default simulation provides an *autobiographical context for experience.* Thus narratives or narrative fragments produced by mental time travel are characterised by *subjective adequacy* rather than truth, accuracy, public acceptability or verification. By subjective adequacy I mean that the story or narrative fragment "fits" with the agent's psychology, not that it is adequate to any public epistemic standard. In fact the opposite is often true. The stories people tell to themselves about themselves in order to cope with experience are often at odds with objective accounts of the causes of their experience.

The reason is that to verify these subjectively adequate narratives the subject has to treat them as hypotheses about the nature of the world or the causes of experience and then engage in the process of confirmation or disconfirmation. To adopt that perspective on the information represented in the story is to be able to *decontextualize* as cognitive psychology puts it. The same representation, the thought of being inhabited by a lizard for example, may be subjectively adequate as an element of a story or fantasy which enables the subject to cope with experience, but straightforwardly falsifiable by decontextualised processes.

Decontextualisation requires activation in specific (dorsolateral) neural structures which are the last to develop, phylogenetically and ontogenetically, and the most cognitively

expensive to deploy. Decontextualisation is not itself a specific form of inference, such as logical reasoning or hypothesis confirmation, reading maps or graphs, but a necessary precondition for such forms of context-independent cognition. This is why the neural correlates of decontextualized processes can be activated in a wide range of tasks ranging from working memory to goal recalibration or abstract symbol manipulation. What these tasks have in common is that they require the subject to represent information in a context-free way. Mercier and Sperber have suggested that decontextualised processing evolved as a way to resolve conflict produced by conflicting subjectively adequate narratives (Mercier and Sperber, 2011a, Mercier and Sperber, 2011b). If you and I disagree about the quickest way to a destination based on our own previous experiences a way to resolve the problem is to look at a map. In other words to represent the problem in context free format, using information not drawn from subjective experience. Obviously one has to represent one's own position on the map to use it but the relationships between locations on the map are independent of the user. In that sense it is a decontextualized form of representation. The same is true for rules of inference used in personal deliberation. One can reason about *oneself* using impersonal, objective rules of inference.

This theory does not suggest that decontextualisation can *only* be used in social/argumentative contexts. Just that it is an adaptation for argument. Clearly it can also play a role in theoretical reasoning unmoored from practical concerns.

In cognitive psychology decontextualisation is often equated with abstract reasoning, standardly tested by problems whose solution requires logical or mathematical reasoning. Because ability to decontextualise is usually tested by ability to perform reasoning tasks, the difference between decontextualised and subjective, autobiographical, forms of thinking is sometimes misleadingly described in terms of a contrast between

(i) Subjective versus rational
(ii) Subjective versus abstract
(iii) Subjective simulations versus beliefs

The difficulty with (i) is that one can reason rationally using objective procedures from a first person perspective or in pursuit of subjective goals. Indeed this is one defining feature of Bayesian reasoning. It allows an individual to confirm hypotheses derived from personal experience. The result may be an idiosyncratic, internally consistent, false, set of subjective beliefs produced by an objectively rational process.

The same is true of (ii). " I am six feet tall therefore I am over five feet tall " is reasoning about myself but here the use of the first person pronoun does not signal that the process expresses a unique subjective perspective on the world. We could substitute a name or description for "I" and the inference would work just as well, logically and psychologically.

Similarly the apparent contrast between simulations and beliefs is only a contrast if beliefs are *of necessity*, symbolic or abstract forms of representation which have nothing to do with simulation. But someone who decides a on a holiday destination by imaginatively simulating alternative holiday experiences and choosing on that basis can plausibly be thought of as forming a belief that one destination is preferable.

The default network and the "essential indexical"

A better way to capture the concept of a subjective perspective on information is to via the concept of "essentially indexical" representations expressed using the first person pronoun.

For example a parent hears a radio bulletin about children killed in a car crash while on a school excursion. As more descriptive information is provided she becomes increasingly worried. What explains that worry except the fact one of those children might be *hers*. The crucial psychological element is the difference between knowing that *a* child "aged seven, with blond hair, wearing a blue sneakers and a red hooded sweatshirt …, …" and "*my* child is seven, is blond, was wearing blue sneakers and a red hooded sweatshirt …, …,"

We can call this difference self-location. It provides the first person perspective on the information conveyed in third-personal propositional form. The same point was made above about maps. One can only use a map if one can locate oneself on it. The point applies quite generally: information represented in a decontextualised form needs to be assimilated to the subjective perspective to motivate.

Self-location is not a cognitively unified phenomenon: it may arise in different ways in different contexts, such as manipulating objects (the motor systems needs to represent the knife in relation to *my* hand to cut the sandwich), or recognising that a description applies to oneself, and at varying levels of explicitness.

Mental time travel involves the explicit simulation of essentially indexical experiences as an essential aspect of high-level cognitive control. This is a consequence of the fact that the default circuits on which it depends are densely connected to emotional and motivational systems. Thus simulations produced by mental time travel are automatically associated with the emotional and motivational processing which provides the spur to action. These wired-in associations are a double-edged sword. They are essential for planning since they transform motivationally inert forms of representation into personally relevant ones. For example when planning my superannuation I can calculate returns on investment mathematically using decontextualised processes. (“It costs X thousand dollars per year to meet living expenses in 2025 based on current projections.”) I can even express the conclusion using the first personal pronoun without it necessarily being motivationally significant (“ I need to save Y thousand dollars per year to produce Z amount”). However those numbers acquire motivational force if I simulate relevant alternative futures: spending my final moments of life writhing on a filthy blanket in a quasi prison for the demented versus drinking champagne on the penthouse terrace of a Dubai resort for hedge fund managers and arms dealers. The point is that these simulations are essentially indexical and help inform the conclusion “ I need Z thousand dollars by 2025”. This is the positive side of essentially indexical simulation. On the negative side, the subjective perspective provided by default simulation can also be

highly misleading. When people are in a good mood and imagine retiring by the Adriatic they happily "remember" summer holidays and honeymoons. The freezing winter gales, empty streets lined by shuttered shops and putrid tideline strewn with sewage tend not to be salient. Mental time travel is a poor source of information in situations for which we have no precedent in experience.

Subjectivity, affective processing and the hub of the default network

These ideas suggest that damage to the default network would affect mental time travel and abilities that depend on it. People unable to perform mental time travel might find that while they are able to reason abstractly about outcomes and alternatives the conclusions they reach are not "felt" in an essentially indexical way because they cannot simulate the relevant experience.

Some initial support for this idea is provided by a distinction made by cognitive psychologists between episodic and autobiographical memory. Episodic memory (for experiences) seems necessary but not sufficient for autobiographical memory. This is shown by the fact that a know/remember task used to distinguish episodic memories from semantic memories for facts is not quite the same task as a know /autobiographical memory task. If I ask you how many chevaliers on the Bayeux Tapestry, or to name the players in the Champions League final you might simulate a visual image but not necessarily your sense of being present. If, however, you evoke a representation of a scene *in which you re-experience yourself as present* then you produce an autobiographical memory. Of course given the constructive nature of memory there is no guarantee that such a memory will accurately represent your experience.

While "pure" episodic memory studies (such as recall of visual scenes) do not activate the ventromedial prefrontal cortex "activations of the ventromedial prefrontal cortex are almost invariably found in *autobiographical* memory studies" (Gilboa, 2004). Gilboa (2004, p1336) suggests that this is because "autobiographical memory relies on a quick intuitive 'feeling of rightness' to monitor the veracity and cohesiveness of retrieved

memories in relation to an activated self-schema". This is consistent with studies showing activity in the ventromedial and related subcortical structures when people make intuitive (that is rapid and semi-automatic judgements about themselves. When people make more decontextualised judgements about themselves dorsolateral prefrontal structures are engaged (Lieberman et al., 2004).

A unifying interpretation of studies of patients with lesions to the ventromedial prefrontal cortex, which document a range of behavioural, decision making, emotional and social problems is that ventromedial patients oscillate between various forms of reflexive stimulus bound cognition and decontextualised processing. They are typically described as impulsive and unable to inhibit automatic emotional responses and routines (Damasio et al., 1990, Damasio et al., 1991, Dunn, 2006, Koenigs and Grafman, 2009). In other words they are more automaton than agent. This is because the *automatic* affective and motivational structures are intact. They are stimulus-driven hostages to current incentives and automatic emotional reactions.

Their ability to detach from a current stimulus to reflect and plan and use that information to control their behaviour is very limited. Yet their semantic knowledge and capacity for decontextualised reasoning is often preserved. This is explained as a consequence of the fact that decontextualised reasoning depends on preserved dorsolateral prefrontal structures. Consequently a ventromedial patient may be able to do a utility calculation about her personal future or the effect on close relatives of her actions but be unable to act on that knowledge. It appears as if semantic knowledge is motivationally inert.

What they have lost is the ability, provided by ventromedial structures to simulate affective and motivational response *in the absence of the stimulus*, while they retain the ability to process information in a decontextualized way. Such results are often used to emphasise the necessary role of affect in deliberation but they also suggest that what those affective responses do is provide the necessary *personal* perspective on information. They make the information "mine" so to speak. Furthermore this diminishment is not just at a time, but over time. These patients, although not amnesic in the strict sense of the

terms have very limited ability for episodic recall or prospection. They do not travel in time because they have no self to project in simulation.

This suggests that disorders in which people feel a diminished sense of self would be characterised by hypoactivity in the ventromedial prefrontal cortex. In a review of the neuropsychological and imaging literature Koenigs and Grafman concluded that "one could conceive of the vmPFC patients selective reduction in depressive symptoms as a secondary effect of *a primary lack of self awareness and self reflection*" (Koenigs and Grafman, 2009). In other words patients with ventromedial damage do not "feel" personally affected when considering even quite distressing events.

It seems that "mine-ness" of experience, *actual or simulated*, is a cognitive achievement mediated by the ventromedial prefrontal cortex. This is consistent with the fact that the ventromedial prefrontal cortex is a hub of the default network. Consequently ventromedial patients not only have diminished affect but seem genuinely unable to project them*selves* in time (Dunn, 2006 ; Damasio, 1991; Bechara, 2005; Gerrans, 2010).

One reason the ventromedial prefrontal cortex is suited to play this role is that it recapitulates at a higher level many of the processing properties of lower level hubs of emotional processing that play the "essential indexical" role for automatic processing. Structures like the amygdala coordinate the automatic information gathering necessary to set up automatic bodily and behavioural responses to salient information represented in perception (Adolphs, 2010, Gottfried et al., 2003, Bechara et al., 1999). Thus the information processed by the amygdala is *intrinsically essentially indexical* since it links perception to action.

Rather than reinvent the cognitive wheel for controlled processing evolution has provided pathways which traffic affective and reward predictive information processed automatically at lower levels to controlled processing coordinated by the ventromedial prefrontal cortex. This allows some forms of explicit processing to inherit the essential indexicality of automatic affective processing. The simulations and symbols manipulated

in these higher levels of processing acquire their motivational force through their communication with lower level affective and motivational systems (Hauser, 2006).

One crude way of thinking of delusion and perhaps some other psychiatric disorders is that delusional or psychiatric patients in general perhaps have anomalies in the way this sense of "mineness" is generated and not enough decontextualized processing. This would suggest that in delusion we would find not hypo but *hyper*activity in the default network accompanied by hypoactivity in circuits required for decontextualized processing.

Anticorrelation between default and decontextualised processes

One thing that seems to be distinctive of delusion and perhaps other psychiatric disorders is that decontextualisation is very difficult for delusional subjects..

One important reason is that the circuits which support default and decontextualized processing are *anticorrelated*. That is to say high levels of activation in one circuit are accompanied by reduced activation in the other. This a consequence of the heavy processing demands made by higher-level cognitive processing. The mind cannot sustain default thinking *and* decontextualised thinking, which are both high level metacognitive processes, simultaneously, so must constantly balance the allocation of resources between them. In delusion the ability to subject a default thought to decontextualized evaluation is compromised.

Experimentally this type of resource allocation is tested in a variety of tasks that require decontextualisation. For example in a working memory task a subject watching letters scrolling across a screen has to determine if sequences have been recently repeated. The interval between repetitions and complexity of sequences can be varied to produce different degrees of difficulty.

In conditions like this neurotypical subjects deactivate the default network and show increased activation in dorsolateral prefrontal areas. In an imaging study using this working memory paradigm " It appears that the brain alternates between activation of the default network, when not engaged in a task and suppression of the default network when engaged in a task" (Whitfield-Gabrielli et. al. p 1280). Tasks requiring the manipulation of impersonal information activate different areas of the prefrontal cortex (dorsolateral) to those ventral parts involved in personal rumination which form part of the default network.

There is emerging evidence that delusional subjects are poor at this kind of inhibition. In the same study in which Whitfield Gabrielli et. al. established that neurotypical subjects inhibit the default network in order to perform an abstract working memory task they compared the performance of subjects with early onset schizophrenia and first degree relatives (Whitfield-Gabrieli et al., 2009). They found *hyperactivity and hyperconnectivity* of the default network in the at rest condition compared to normal subjects. They also found that the schizophrenic and first degree relatives showed reduced deactivation of the default network in the working memory task and reduced anticorrelations between medial prefrontal cortex (part of the default network) and dorsolateral prefrontal cortex (activated for the impersonal working memory task).

These results are consistent with the idea that delusional subjects (at least the schizophrenic ones tested here) devote excessive cognitive resources to constructing default simulations.

However the problem is not just a failure to inhibit mental time travel which is otherwise "normal". A study by D'Argembeau and van der Linden tested the ability of schizophrenics with positive symptoms (delusions and hallucinations) on mental time travel tasks which asked them to evoke temporally distant episodes. " Results showed that patients recalled fewer specific past events than did healthy controls and were even more impaired in generating specific future events "
(D'Argembeau et al.) (p247).

A restricted focus on current experience is consistent with some of the observed phenomenological facts about schizophrenia as well as some of the cognitive studies. Phenomenologically the psychotic episodes in schizophrenia are often preceded by a prodrome in which experience takes on an altered feeling of significance. The “delusional mood” as Jaspers and other called it. Another feature of the prodrome is concurrent anxiety or depressive affect which also, via interactions between affective and salience systems, have the characteristic effect of monopolizing cognitive resources (Heinz and Schlagenhauf, 2010). Thus the role for default circuitry suggested here is consistent with the observation that schizophrenic subjects in particular are preoccupied with subjective experience. Not only that but they find it very difficult to escape the biasing effect of that current experience on their default thought processes. All their thoughts are essentially indexical. Not in the sense that they automatically produce actions but in the sense that they amplify the “ here and now for me ” at the expense of contextually remote information.

A mundane example

The mind is always poised to allocate resources according to the nature and relative importance of problems. For example while driving to the airport my mind is largely on automatic, controlling the car in response to perceptual input. In a series of studies Mason et all have shown that “several regions in the de- fault network exhibited greater activation during highly practiced tasks that allow for mind wandering, including the medial prefrontal cortex, the anterior cingulate cortex, the posterior cingulate cortex, and the precuneus, among several (Mason et al., 2007). They then propose that mind wandering is ‘‘a psychological baseline that emerges when the brain is otherwise unoccupied, supported by activity in a default network of cortical regions’’ (Mason et al., 2007, p. 394).

In the absence of any pressing problem to deliberate about my DMN reverts to mind wandering mode. As I daydream narrative elements associate in scenarios, assembled by causation, contiguity and resemblance. I anticipate a happy reunion with my old friend

the Regius Professor of Gender Studies. I remember the Regius Professor has just tweeted about an experiment that contradicts my ideas about delusion. Enough default thinking! Fortunately the road is clear and driving easy so I can continue to delegate the driving to my automatic processes. I inhibit my default daydreaming and engage decontextualised processes to reconstruct my arguments. My argument repaired after twenty minutes careful thought activity in my decontextualised systems subsides and activity in my DMN increases. I resume daydreaming.

Suddenly I notice that I have missed my turnoff. This is an instance of a low level perceptual prediction being violated, producing surprisal, referred to controlled cognition. I now engage in supervised mental time travel as I mentally rehearse alternative journeys then switch to switch to more decontextualized mode as I recalculate schedules.

This interplay between perceptual capture and top down control, directed and undirected default thinking and decontextualised thought is the human condition, which ceaseth only in death. As we shall see below default thinking, entirely undirected, continues in sleep, in the form of dreaming. Perhaps in the middle of the night I will awake from a nightmare, like Bertrand Russell's about Anselm, shouting "Great God in Boots. The Regius Professor of Gender Studies is right after all!"

The default network never sleeps. In fact it is most active during sleep because it is entirely unsupervised.

DREAMING DEFAULT THINKING AND DELUSION

In this chapter we pursue the idea that the often-remarked, but not well explained, similarity between dreaming and delusion results from the fact that they are both states characterised by activity in the default system unsupervised by decontextualised processes. We give a mechanistic and a cognitive explanation of similarities and dissimilarities between delusions and dreams. The mechanistic explanation adapts the AIM (Activation, Information, Mode) model of dreams and delusions developed by Hobson and collaborators. The cognitive explanation shows how automatic and controlled processes degrade in different ways when unsupervised in virtue of differences in cognitive architecture.

That explanation distinguishes feature and context binding. Binding here refers to the integration of elements of a representation. Feature binding refers to the construction of a integrated perceptual or quasi-perceptual representation and contextual binding to the organisation of such representations into a coherent metacognitive structure such as a narrative or theoretical explanation. Feature binding is a relatively modular process implemented in localised neural circuits whose processing properties are fairy rigid. In general feature binding is a task performed by weight based processing systems which

automatically assemble representational elements according to procedures standardised by evolution or in development. Because these processes have fairly rigid architectures they tend not to disintegrate unless their neural substrates are damaged. Context binding is intrinsically a more unstable process because it is by its nature flexible and open ended. It requires active maintenance of transient distributed circuits and constant top down supervision.

One form of context binding relevant to delusions is of course the binding of simulations into the narratives and narrative fragments of mental time travel.

Delusions are a form of context binding in which subjects produce a default thought subjectively adequate to experience. That thought is not the result of an attempt to confirm an empirical hypothesis. It is an untested (because unsupervised) default thought. We develop these ideas by discussing an example of delusion-like phenomenon in dreams: the experience of *hyperfamiliarity*, and suggest that it results from a similar-feature biding anomaly to one which occurs in a waking state in the Fregoli delusion. In dreams however this anomaly does not lead to delusion because in dreaming the default network is not configured to produce a narrative context for such anomalous experiences.

Dreaming and the Default Mode Network

The default system evolved as the first stage in the transformation of automatism to agency. It allows humans to rehearse alternatives before committing to action. These rehearsals typically take a narrative form. The production and consumption of these narratives is in turn supervised by a higher level of *decontextualised* cognitive processing. This supervision can take the form of testing narrative elements for consistency and veridicality or evaluation of competing narratives for accuracy or utility.

When dorsolateral systems that represent high-level goals and evaluate default narratives against them are inactive the default network reverts to its default state: the production of subjectively adequate narratives. When there is no goal or end point for for the narrative

it tends to degraded into the random association of default representations. This is sometimes put in terms of the absence of attentional control in default thinking. A simulation process that starts with a specific goal may meander as top down attentional control wanes.

It is for this reason that dreaming represents a pure state of unsupervised default processing. As well as the absence of top down control the default processing in dreaming is characterised by absence of organising input from the sensory periphery which leaves the default network entirely at the mercy of endogenous activation by subcortical inputs.

> dreams can be seen as a unique and more fully developed form of mind wandering, and therefore as the quintessential cognitive simulation. They are the quintessential cognitive simulation not only because they have elaborate story lines that are often enacted with exquisite sensory, motor, and cognitive involvement, with some dreams unfolding over a period of several minutes to half an hour or more. There is also the striking fact that they are usually experienced as real while they are happening. (Domhoff, 2011)

This is one reason that people have drawn a parallel between dreaming and delusion.. John Nash, for example, said of his delusional states "Its kind of like a dream. In a dream it's typical not to be rational". While this is correct it slightly mischaracterises the phenomenon. Some dream reports include episodes of coherent thought. However such episodes are transient and not coherent with other episodes of the dream. Equally some delusions and dream episodes can include episodes of rational transition between thoughts that are bizarre and false. Paranoid delusions for example form coherent and sometimes logical narratives given the thoughts that trigger them. For this reason the similarity between dreams and delusions is sometimes put in terms of absent or compromised "reality testing" (Bentall, 2004). "Reality testing" is a slight term of art and clearly can include evaluating propositions for truth or experience for veridicality. But

John Nash, while remaining capable of high-level mathematical reasoning never thought to question the reality of his delusion that the Pope and CIA were part of a conspiracy against him. So the term really refers the suspension or absence, partial or complete, of the ability to detect and resolve inconsistency between the information represented in an episode of thought and the background knowledge about the world, normally available to be brought to bear in alert waking cognition.

One reason for taking the cognitive approach to the phenomenon is that it avoids a controversy over the exact nature of "reality testing". Sometimes it is tested in paradigms which, in effect, ask the subject to reason about possible causes of experience. A version of this paradigm is employed in source-monitoring experiments which compare the ability of delusional and non-delusional subjects to estimate the probability that an experience has endogenous or exogenous origin (Mitchell and Johnson, 2000, Moritz et al., 2003). Naturally these paradigms are congenial to the Bayesian model, which treats source monitoring as the detection of prediction error.

While remaining neutral about the implementation of Bayesian algorithms we can note that there is clear evidence for

(i) A role for dorsolateral circuitry in the supervision of default thinking
(ii) Hypoactivity or lesion of that circuitry in conditions such as delusion.
(iii) Deactivation of that circuitry in dreams
(iv) Activity or hyperactivity of the default circuitry in both delusions and dreams

Two well known dream theorists explain the resultant phenomenonology as follows, also providing an explanation of the personal level consequences.

> much of wake-learned knowledge is not accessible or even applicable to the largely self contained dream narrative. This alleged inability to exchange information is buttressed by experiments which have shown that there is a decrease in DLPFC with orbital prefrontal exchange specifically during rapid eye

> movement (REM) stage when dreaming is pronounced and, a decrease in information exchange of DLPFC with parts of the occipital, parietal and temporal lobes .

Kahn et. al. make a similar point:

> When the DLPFC is in poor communication with these areas as in REM sleep, the ability to perform logical inference, to recall accurately and to discern whether a premise is fact or fiction may very well be compromised (Kahn et al., 2002).

The most extreme (and congenial to the thesis developed here) attempt is by Gottesman who explicitly situates his neurobiological account of dreaming and schizophrenic psychosis in a framework which treats the mind as the mind as a control hierarchy with decontextualised supervision at the top. He explains the properties of the dream state in terms of the absence of top down control due to the deactivation of the dorsolaterally-regulated supervisory systems, in combination with the absence of perceptual and sensory input.

> the dorsolateral prefrontal deactivation observed both during REM sleep and in schizophrenia seems to suppress or decrease its own functions, including the loss or decrease of reflectiveness, and at the same time disinhibits older subcortical structures and corresponding functions, with the exaggeration of accumbens' and amygdala nuclei's own processes: in our case, the appearance of hallucinations, delusions, bizarre thought processes, and affective disturbances. (Gottesmann, 2006).

Thus there is consensus that what we have called decontextualised supervision is absent in dreams due to the deactivation of the dorsolateral prefrontal circuitry on which it depends. ''REM sleep may constitute a state of generalized brain activity with the specific exclusion of executive systems that normally participate in the highest order analysis and integration of neural information.'' (Braun et. al.1997 p1190)

We first give a version of the neurochemical account of dreaming (the AIM model) and delusion before marrying it to the binding theory of cognitive coherence. Although the AIM model has generated controversial interpretations (for example of the relationship between REM sleep, dream experience and activity in specific brainstem structures) we rely here only on those uncontroversial theses about the mechanisms involved in context binding. We then discuss a concrete example using the Fregoli delusion of the way in which the account can work to explain similarities between dreams and delusions.

The AIM model

The AIM model (Activation, Information, Mode) model of dreaming is based on the fact that changes in patterns of activation across the brain change the flow of information within and between components of the control hierarchy. The intrinsic cognitive properties of these components are preserved through transitions from mode to mode. What changes are the interactions between these components. Different modes (Rapid Eye Movement sleep, Non Rapid Eye Movement sleep, alert waking, daydreaming) reflect differences in cognitive coherence or the way in which the components of the hierarchy are integrated.

The AIM model also points out that transitions between modes of waking cognition (from alert focussed attention to default mode for example) are comparatively subtle instances of larger scale modal transitions of which the most important is sleeping and waking. Within sleep the important modal transition is between Non Rapid Eye Movement (NREM) and Rapid Eye Movement (REM) sleep. The table below integrates the AIM model with the discussion of controlled and automatic processing in previous chapters.

Behavioural mode	**Wake**	**NREM**	**REM**
Cognitive modes **Automatic Processing** **Controlled processing** Default Thought Decontextualised Thought	Acquisition and manipulation of information Perceptual Sensorimotor Loops Affective response Metacognition Subjective context for salient experience Detection of inconsistency. Reality Testing	Iteration of Information Rehearsal of standard patterns of behaviour and familiar scenarios	Integration of Information
Phenomenology: Perception	External, vivid Continuous,	Dull or absent Episodic	Internal, vivid

Movement Thought	voluntary Logical Coherent Progressive	Involuntary Logical Coherent Perseverative	Commanded but Inhibited Incoherent Associative
Neurochemistry 5HT (serotonin), NA(Noradrenaline) Cholinergic DA dopamine	 High Low Salience within a functioning control hierarchy	 Decreasing at end of cycle Increasing at end of cycle Tonic. stabilise scenarios	 Low High Phasic. turnover elements of default system

These transitions between modes are essentially a consequence of the balance between cholinergic (acetylcholine) and aminergic (5 Hydroxytryptamine /serotonin, norepinephrine and dopamine) regulation of the brain by reticular activating systems which project throughout the brain from the brainstem.

Very roughly aminergic regulation is required for wakeful exploration of the environment and goal-directed behaviour and cognition. Amines assist the prefrontal cortical structures to communicate with posterior ones and to build and actively maintain transient distributed networks necessary for controlled processing. During active waking characterised by slow wave firing patterns across the brain maintained by tonic levels of 5HT in particular the noradrenergic, sero-toninergic, and cholinergic systems are firing together. During slow-wave (NREM) sleep, these three systems reduce their discharge,

while during REM (rapid eye movement) sleep, the noradrenergic and 5HT systems shut off completely, as the cholinergic system resumes its firing**.** However the dopamine system projecting from the *nucleus accumbens* remains active even during REM sleep.

Serotonin seems to enable the construction of stable patterns of activation across widely distributed neural circuits, enabling global integration of different systems for particular tasks involved in wakeful exploratory activity. If 5HT levels are reduced in rats their foraging, sexual and social activity is reduced (Allman, 1999). Even in animals with almost no brain, serotonin neurons are involved in orienting animal towards nutrient sources and controlling digestion. This suggests that the regular delivery of serotonin across the brain keeps the organism in state of wakeful exploration, coordinating perceptual motor and cognitive activity to enable life-supporting activities. Of course what constitutes such an activity, and the neural circuitry which supports it varies from organism to organism so a monolithic *representational* interpretation of the role of serotonin across all neural circuitry is impossible. This is why Hobson introduced the notion of cognitive mode. It describes the global functional integration and coordination of cognitive subsystems when the mind is in exploratory mode.

Acetylcholine plays a role in maintaining patterns of activation across localised assemblies rather than the global patterns maintained by the serotonin system. It is thus an antagonist to 5HT, adaptively disrupting stable global patterns of activity and detaching local assemblies from global integration. One feature of cholinergic regulation is that when it increases in wakeful stressful episodes prefrontal activation is reduced. High levels thus produce a reversion to automatic mode. This has consequences for the understanding of many developmental disorders and stress disorders manifest in episodes of disinhibition. As with other neurotransmitters a one to one correlation with a cognitive function has not been proposed. Rather the balance with other neurotransmitters and the density and type of receptors in areas to which it projects determine how it influences cognition. However we can note that when choline levels are high relative to amines they produce lack of prefrontally-based executive supervision and reversion to automaticity.

The norepinephrine system pays a crucial role in poising the system for defensive action ("flight or fight"). Circuits it innervates control alertness vigilance and agonistic behaviour releasing appropriate hormones and priming visceral and somatic systems for rapid response.

Finally as we saw previously the dopamine system is a crucial part of the salience system. It enables an organism to target attention and cognitive effort on relevant stimuli by accentuating activation in circuits which represent salient information.

Of course at any given moment the effect of any neurotransmitter is highly context dependent. It depends on density and type of receptors, mode of delivery and current activation level and neurochemical state of the circuit it innervates and the representational architecture of those circuits. No monolithic interpretation of the role of a neurotransmitter is possible for this reason. Dopamine for example plays roles in learning and memory as well as motor control (evidenced by Parkinson's symptoms caused by low levels of dopamine). Serotonin causes blood vessels to contract as well as influencing a wide variety of prefrontally based functions.

Nonetheless the global state of the brain (sleep, wake, explore, withdraw, automatic, controlled) is regulated by fundamental neurochemistry because that chemistry synchronises global activation patterns at different ranges and time scales.

When the balance of cholines and amines distributed across brain regions by the reticular activating systems changes in favour of cholines, prefrontal activity, and with it the capacity for metacognitive responses to experience, subsides. Simultaneously, motor expression is inhibited and early stages of perceptual and sensory processing are shut down. Alert wakefulness and REM sleep are the ends of a cycle with NREM sleep constituting an intermediate stage, neurochemically and cognitively (see table above). In NREM sleep there is no perceptual input and metacognitive supervision is reduced. Volitional control is largely absent as a consequence of prefrontal deactivation. Consequently standard routines or associative repertoires tend to be replayed, often

accompanied by negative affect since the emotional systems remain active but there is no top down integration or supervision of these routines.

When serotonin is at its lowest level and the brain is cholinergically regulated automatic processes continue without being organised externally by the environment through perceptual inputs or internally under volitional control. This is REM sleep in which the mind is entirely at the mercy of unsupervised perceptual, sensory and emotional processing.

In dreams some cognitive processes, such as late states of perceptual processing, remain relatively intact, producing the characteristic stream of imagery. Others, such as logical argumentation, volitional control and planning are absent or reduced. Thus in REM dreams we experience cognitive fragments such as images and sensations juxtaposed incongruously in vignettes and scenarios rather than coherently organised in narratives or explanations (Schwartz, 2002; Schwartz, 2002a; Manni, 2005; Solms, 2007 ; Roehrenbach, 1995 ;Hobson, 1999; Revonsuo, 2002; Dawson, 2011; Domhoff, 2005 ; Manni, 2005; Revonsuo, 2006, Revonsuo and Salmivalli, 1995, Revonsuo and Tarkko, 2002).

Feature Binding and the Fregoli Delusion

The phenomenon of hyperfamiliarity experiences seems common to both dreams and some delusions of misidentification. In delusions of misidentification people report seeing “doubles”, imposters, people in disguise, people changing appearance and identity (Breen et al., 2000, de Pauw and Szulecka, 1988, Ellis et al., 1994, Spier, 1992, Stone and Young, 1997). The phenomenology of these delusions can be explained in terms of abnormalities of feature binding, combined with abnormalities in context binding. The abnormality of feature binding produces a representation in which elements normally bound together such as face, name, autonomic response to a familiar person, and identity may dissociate . We mentioned one such delusion, the Capgras delusion, maintained for a whole family, in the introduction. In another delusion of misidentification, the Fregoli

delusion patients report being followed by a familiar person in disguise (Courbon and Fail, 1927, de Pauw et al., 1987, Ellis and Szulecka, 1996, Ellis et al., 1994, Eva and Perry, 1993, Joseph and O'Leary, 1987, Wolfe and McKenzie, 1994, Wright et al., 1993). An explanation based on the cognitive architecture of face recognition is that the patient sees an unfamiliar person but has a strong affective response characteristic of seeing a familiar person. This incongruity is produced within dedicated circuitry that recognises faces and matches them to identifying information and initiates appropriate affective responses. The patient then has the experience of seeing a stranger but feeling as if the stranger is familiar to her. The patient then deals with this incongruity by producing the delusion " I am being followed by a familiar person in disguise".

Recently cognitive theorists have taken a unified approach to these phenomena noting that these Fregoli-like phenomena in which strangers are "felt" to be familiar also occur in dreams. These cases of "hyperfamiliarity" in delusion and in dreaming are explained in terms of similar abnormalities of feature binding. The differences are explained in terms of characteristic differences in context binding in each state.

In order to explain feature binding these theorists use a cognitive theory of face recognition developed to explain disorders of face recognition known as prosopagnosia and extended initially to explain the Capgras delusion (that a familiar person has been replaced by an imposter or lookalike) and then other delusions of misidentification such as Fregoli.

In prosopagnosia comes in different forms but the relevant one for us is one in which a subject can recognise a face *qua* face *and* recognise facial features but is unable to determine whether a seen face is familiar to her. Such patients will produce identical responses when presented with both novel and previously seen faces (Bauer, 1984, Bauer, 1986, Bruyer, 1991, Ellis et al., 1993b, Landis et al., 1986, Sergent and Signoret, 1992, Shraberg and Weitzel, 1979). In associative prosopagnosia the subject has no explicit (overt as it is called in the face recognition literature) awareness that she has seen a face before. Nonetheless in some cases lack of explicit recognition is accompanied by

preserved implicit recognition (called covert recognition in the face recognition literature). Covert recognition comes in two forms: the first I will call behavioural covert recognition (BCR). The second is skin conductance response (SCR). These covert forms usually co-occur but it is useful to distinguish them.

In the behavioural case, covert recognition is evidenced by things like Stroop-type interference effects, response time priming and by a paradigm called true face name relearning. Interference effects were shown in a task in which a prospagnosic patient was asked to classify names presented in a list (e.g politician or actor). His performance was affected by the presentation of a distractor face from the wrong semantic category (e.g. George Cluny presented alongside David Cameron) (De Haan et. al., 1987). The point of these interference cases is that the distractor face could only interfere with performance if identified. In response time priming prosopagnosic patients are asked to judge familiarity of names, which typically presents no problem (their problem is with faces not names). However their response time is faster if they have previously been presented with the named face. This should not be the case if they are genuinely unable to recognise faces. In true face relearning patients are presented with pairs of faces and names. In half the cases the names are incorrect. The patients are then re-presented with the pairs after an interval and asked to name the face. The fact that their performance is better in the "true" case is evidence of covert or implicit recognition (Schweinberger and Burton, 2003, Bruyer, 1991)Sperber and Spinnler, 2003, Bruyer et. al. 1993).

An example of covert SCR was provided by Bauer who tested the galvanic skin responses of an overtly prosopagnosic patient, LF, to photos of familiar and unfamiliar faces. LF could not spontaneously name a single face and guessed names at chance level from a list of five, but covertly responded to more than 60% of familiar faces (Bauer, 1984, Bauer, 1986)

It is possible that SCR is simply a way of testing for BCR, that is to say that it is another indication of activation in the same neural system responsible for true face name

relearning and response time priming. In fact this is one way to interpret Bauer's early work in the area.

However, there is some evidence to suggest that for the processing of faces, SCR and BCR depend on processing in distinct pathways. BCR is indicative of early stages of processing in the pathways culminating in face recognition. SCR is indicative of the subject's autonomic response to the face consequent on "early" activation in this pathway.

Consequently as Young puts it, "it is inadequate to think of it [prosopagnosia] as simply involving loss of recognitional mechanisms. Instead, at least *some degree of recognition* does take place. What has been lost is *awareness of recognition*." (quoted by Schweinberger and Burton , 2003, p284. My italics).

A model which explains how this can be the case is now a basis for explanation of the Capgras delusion

INSERT FIGURE 1 ABOUT HERE

LEGEND: Basic Architecture of the face processing system: from perception to identification.

The standard account of delusions of misidentification is that autonomic response is "mismatched" to a seen face. In Capgras delusion the autonomic response to a familiar is absent. This makes Capgras delusion in effect a double dissociation with prosopagnosia with preserved SCR. This is shown in the above figure above by the locations of the relevant lesions. In Capgras the lesion is to pathways connecting the face recognition units and the amygdala. In prosopagnosia the lesion is to pathways leading from the FRU to later more explicit forms of recognition.

In the Capgras delusion the patient sees a familiar person but does not have the characteristic autonomic, affective response. The delusion is an attempt at context

binding which accounts for the familiar appearance and the absence of affective response. Doubles or imposters appear familiar but do not evoke affective responses.

The same model can also explain the Fregoli delusion as cases where perception of strangers evokes an autonomic response characteristic of familiars due to hyperactivity in pathways which trigger the autonomic response to familiars. A theoretically economical explanation of the relationship between Fregoli and Capgras would equate them to pathological instances of *deja* and *jamais vu* for faces instead of places. *Déjà vu* for place involves a feeling of familiarity for a place that is not recognised as familiar. *Jamais vu* is the reverse: A feeling of unfamiliarity simultaneous with recognition that the place is in fact familiar. In fact this analogy has been explicitly drawn a number of times (O'Connor and Moulin, 2010).

The same model has been invoked to explain dream experiences of hyperfamiliarity. Revuonso and Tarrko (2002) analysed 592 dream reports which contained bizarre dream characters (people who appear in dreams). They divide bizarreness into two types which correspond to relational (e.g. lives in UK) and intrinsic (facial appearance, familiarity and identity) properties of the character. Among the latter, which they hypothesise is essentially a consequence of feature-binding in the face recognition system, they include " the emotional component signalling the familiarity or unfamiliarity of the person". Interestingly, and in accordance with the hypothesis that the modularity of the cognitive process which produces it corresponds to the degree of representational disintegration in dreams, relational bizarreness was more frequent than intrinsic. Thus characters encountered in unusual or impossible settings or behaving uncharacteristically are more frequently reported than characters with unusual appearance. The latter account for only 2.2% of bizarre reports while the former account for more than fifty percent. *Bizarreness of familiarity* accounts for 5.7%.

This is consistent with the standard model in which face perception and recognition is highly modular but the link between a face and semantic knowledge about a person is not. Thus mismatches between face and semantic knowledge are quite possible if the relevant

circuits are not linked by perceptually-driven synchronisation of activity. The link between face recognition and autonomic response is less modularised than face recognition *per se*, but far more modular than the linking of semantic knowledge and familiar faces. Consequently, as the authors note, inappropriate feelings of familiarity are one of the most common reports of internal bizarreness. “we may interpret the abundance of pseudo-familiar persons in dreams as over-activation of face recognition units: they fire even when the face percept does not correspond to any of the descriptions in the store. This creates the inappropriate feeling of familiarity” (Revonsuo and Tarkko, 2002)

If the standard model is taken as a guide it may be more accurate to suggest that a feeling of familiarity is *a consequence* of activity in the FRU. That activity in the face recognition unit propagates to the amygdala which initiates an autonomic response which is “felt” as feeling of familiarity.

The authors explain erroneous familiarity terms of the relationship between the face recognition unit and the autonomic response. If an unfamiliar face is mistakenly represented as familiar an autonomic response is generated. A similar effect could be produced by endogenous activation of subcortical structures which generate autonomic response. Since limbic systems are active in dreams but not driven by perceptual inputs this is another possible cause of inappropriate feelings of familiarity. In either case the subject could have an autonomic response to a face which is not in fact familiar. Schwartz and Maquet (2002) note in their discussion of dream experience that it resembles the phenomenology of other neuropsychological disorder such as achromatopsia, polyopsia and microtopsia. All these cases can be explained in terms of abnormal patterns of activity produced when later processing in the ventral visual processing stream is driven by sub cortical afferents rather than its usual inputs from early vision. The consequence is abnormalities of feature binding in systems which, because automatic, retain a high degree of cognitive coherence. Rather than disintegrate completely perceptual processes assemble their components differently.

Thus hyperfamiliarity is a relatively frequent occurrence in dreams. As we saw in waking states the delusion that one is being followed by a familiar person in disguise explains why feelings of familiarity are evoked by someone who is not recognised as familiar. In dreams of course no such explanation is forthcoming.

Slighly different Fregoli-like phenomena have been remarked on by Schwartz and Maquet. (Schwartz and Maquet, 2002, Roehrenbach and Landis, 1995). They use the following examples .“ I had a talk with your colleague but she looked differently, much younger, like someone I went to school with, perhaps a 13 year old girl”. In another case a subject reported “ I recognise A's sister. I am surprised by her beard. She looks much more like a man, with a beard and a big nose” Schwarz and Maquet describe these as “Fregoli-like” phenomenona generated by activation in the facial fusiform area. Once again it is plausible to see these cases as candidates for delusional rationalisation: if the sister was disguised as a man it would explain her masculine appearance.

An interesting feature of this type of Fregoli-like report is that the mismatch here is between facial appearance and identity (“ your sister”) rather than facial appearance and autonomic response. In this sense the phenomenon perhaps resembles the phenomenon described by Revuonso and Tarrko as appearing infrequently in dreams “ cases of imposter relatives; person we have never met in the real world but who we accept as our ‘sisters’ ‘brothers’ or ‘cousins’ or ‘uncles’ in the dream.”

Such cases are in some ways the converse of the converse of the Capgras delusion in which a familiar person is identified as an imposter. Very interestingly there are almost no reports of Capgras-like phenomena in dreams. This suggests to dream theorists that the patterns of activity in the facial fusiform area which produce the Capgras delusion do not arise in dreams: “ thus we should expect the pattern of brain activity to be different in normal REM sleep and Capgras patients”.

One likely reason is that identity and appearance are more rigidly bound than affective response to processing within the face recognition system. Thus mismatches between

downstream affective response and appearance are more likely since the link is intrinsically more flexible.

The hypothesis of dream theorists about these phenomena of hyperfamiliarity or mismatch between appearance and identify in dreams is that they depend on similarities between patterns of brain activity in REM sleep and the delusional waking state. Precisely why Fregoli-like phenomena but not Capgras phenomena arise must depend on the nuances of circuitry and physiology.

Context binding in dreams and delusions

At one point in his discussion of dreaming. Alan Hobson says that the ventromedial prefrontal cortex vainly tries to "impose a plot" on the flux of experience but is doomed to fail because it is not in communication which other structures such as DLPFC required to provide coherence and consistency.

In REM sleep the dorsolateral prefrontal cortex "remains conspicuously deactivated" (Hobson p691) and the mind is dominated by highly salient imagery and emotional sensations. The resources to respond to inconsistencies and gaps in the narrative or to represent an overall goal for the narrative are absent due to dorsolateral deactivation, and the hyperassociation of salient imagery continues unbroken.

This selective recruitment of dorsolateral areas for decontextualised, and ventromedial for essentially indexical, thought is a feature of waking cognition replayed in dreaming. In studies of dreaming in prefrontal lesion patients (many of whom experience intense and disturbing dreams) Mark Solms found that dreams were suppressed in ventromedial patients and patients with lesions to sensory association areas which process inputs from "late" perceptual and sensory areas. In other words those components of the default network which produce and associate perceptual and sensory imagery. In contrast patients with dorsolateral lesions who exhibit the characteristic deficits in decontextualised problem solving had normal dreams.

Thus the similarity between dream experience and delusion can be explained by the fact that they share important properties of default thinking, although delusions occur in a waking state. Namely they involve the endogenous activation of components of a network which associates a stream of personally relevant experiences in the absence of actual perceptual input or supervision by decontextualised thought. This captures Hobson's idea that the ventromedial prefrontal cortex "tries to impose a plot". However there is no final reel in dreaming. It ends when the aminergic systems reactivate and their antagonists, the cholinergic systems, reduce their firing rates, according to their own diurnal imperatives.

Dorsolateral deactivation in dreams and delusions

We have already seen that delusions are also characterised by reduced or absent activity *relative to the default network* in experiments in which schizophrenic patients did not produce the normal " task induced deactivation" of the default network in a working memory task.

Thus relative hyopactivity in DLPC is a feature of some delusions. An even more dramatic example of the association between reduced DLPFC activity and delusion especially relevant to the cases discussed here is provided by cases of lesion to the DLPFC associated with delusions of misidentification. Papageorgiou et. al. studied nine patients with delusions of misidentification (Capgras, Fregoli, Intermetamorphosis with some patients suffering more than one of these delusions) (Papageorgiou et al., 2005, Papageorgiou et al., 2003). The 2005 study was an event related potential (ERP) focusing on the P300 ERP component. The delusional patients group showed significant reductions in P300 amplitude in the right frontal hemisphere compared to controls. If we accept the standard hypothesis that abnormal feature binding in the face recognition system is part of the causal chain leading to delusions of misidentification then these cases are instances of abnormal feature binding occurring in a mind with reduced DLPFC function.

The conclusion that hypoactivity of the right hemisphere is involved in delusion is consistent other neuropsychological studies. For example in a group of 33 patients with Alzheimer's disease, of whom 18 had a content specific delusion concerning place person or object, Single Positron Emission Topography revealed hypoperfusion in the right frontal hemisphere in the delusional groups compared to the 15 non delusional sufferers.

Are delusions dreams?

Delusions are not dreams, but as in dreams the balance of activity between default and decontextualized processing has changed. In dreams the default system churns out highly salient simulations triggered by activity in automatic feature binding systems unmoored from the environment. The incongruities and inconsistencies of both feature and context binding are not detected or resolved. Thus narrative incoherence as well as (a lesser degree of) feature binding incoherence is characteristic of REM dreams.

In delusion context binding is more intact. Subjectively adequate coherent default narratives, or elements of them, are triggered by salient experiences. What seems to be missing in delusion is the ability to subject that narrative to decontextualized evaluation.

In the next chapter we turn to the question of how to characterise that lack of supervision.

THE SECOND FACTOR

Most cognitive theories of delusion converge on the idea that perceptual or sensory processing anomalies alone are insufficient to lead to delusion. There must be an additional cognitive process involved in generating delusion.

This chapter discusses the most influential account of this "second factor". That account is a *doxastic* one which treats delusions as beliefs about the causes of experience. The doxastic account (reflected in the clinical characterisation of the DSM) suggests that delusion consists in making bad inferences on the basis of sensory information. However the idea that delusional patients suffer from a pervasive reasoning deficit is unpersuasive. Delusions usually coexist with intact capacities for reasoning on non-delusional topics. Not only that but delusions often seem to be situated in chains of reasoning which, while incorrect, are intelligible. Consequently the most influential doxastic accounts explain delusions as the outcome of reasoning biases rather than reasoning deficits.

Doxastic accounts are drawn from philosophical accounts of belief fixation in which experience is treated as providing evidence for empirical theories. In fact of course belief fixation does not operate directly on experience but on candidate beliefs consistent with experience. Experience prompts or triggers a candidate belief that is then confirmed or

disconfirmed by processes of belief fixation. So there is a crucial intermediate step between experience and the fixation of belief that consists in the production of a candidate belief that purportedly explains the experience.

For example in delusions of misidentification such as the Capgras, automatic feature binding processes produce a mismatch between appearance and identifying information. The initial thought "my wife has been replaced by an imposter" then occurs. Doxastic theorists treat such thoughts as candidate beliefs that are then evaluated for consistency with evidence and background belief by the process of belief fixation. Delusions are explained in terms of the functioning of the belief fixation system. In the case of delusions of misidentification something apparently goes amiss with this process. Instead of rejecting the candidate belief and adopting a more plausible hypothesis the subject accepts it.

In conceiving of things this way the doxastic theorist aligns delusional and normal belief fixation with philosophical reconstruction of scientific theorising. This allows work on biases, on rationality, probability and logic to be integrated with cognitive theories proposed as the basis for integrative explanation. In order to do this the doxastic theorist

(i) Has to treat the initial thought as an abductive hypothesis. That is as a candidate belief generated to explain the sensory information.
(ii) Has to explain why that hypothesis is not rejected or disconfirmed. It is at this point that notions of reasoning bias or deficit are introduced.
(iii) Has to interpret the hypoactivity or absence of activation in dorsolateral regions in delusion as introducing a bias or deficit into the process of belief fixation.

In the next chapter I will argue that (i) that the thought triggered by sensory processing is not a candidate hypothesis generated as part of a process of empirical theorising. It is produced by the default system engaged in the process of providing a subjectively adequate response to salient information. This explains why (ii) it is not rejected on

grounds of inconsistency. (iii) The reduced dorsolateral activity makes it difficult or impossible for the patient to decontextualize and treat this narrative fragment as an empirical hypothesis. Consequently the delusional subject remains hostage to this subjectively adequate default thought.

In the meantime I turn to doxastic theories

Doxastic theories and the second factor.

According to Brendan Maher

delusions are derived by cognitive activity that is essentially indistinguishable from that employed by non patients, by scientists, and by people generally… in brief, then, a delusion is *an hypothesis* designed to explain *unusual perceptual phenomena* and developed through the operation of *normal cognitive processes*" (Maher, 1988)

Maher's idea suggests for example that in delusions of misidentification anomalous sensory processing such as misbinding in the face recognition system is sufficient to generate delusion. No additional cognitive defect or anomaly is involved. It is for this reason that his view has become known as a one-stage or one-factor explanation of delusion. One factor accounts have also been called empiricist accounts because they imply that explanation of the relevant experience exhausts the explanation of delusion (Bayne and Pacherie, 2004, Hohwy, 2004). In the case of delusions of misidentification for example the empiricist theory implies that once we explain the role of anomalous feature binding in the face recognition system in producing an experience of mismatch between appearance and affective response we have explained the consequent delusion.

At first sight it seems that several pieces of related evidence torpedo Maher's account. For example Coltheart (2003) and others note that lesion of the FFA which produces misidentification experiences is *insufficient* to produce delusions of misidentification. Some hypoactivity of dorsolateral prefrontal circuitry also seems to be required. The

most dramatic examples come from lesion studies of nine patients with delusions of misidentification discussed in the previous chapters. (Papageorgiou et al., 2003, Papageorgiou et al., 2005).

A related finding discussed previously is the reduced anticorrelation between default circuitry and dorsolateral circuitry in delusional patients. This also suggests that delusions involve hyperactivity of default circuitry at the expense of dorsolateral (Whitfield-Gabrieli et al., 2009, Pomarol-Clotet et al., 2008).

It does seem that reduced or absent activation in prefrontal dorsolateral circuitry, absolute or *relative to activity in default circuitry*, is involved in delusion. Thus to defend a one factor theory in these cases seems to involve the claim that normal cognitive processes are unaffected by hypoactivity or lesion of the right dorsolateral prefrontal areas. Maher's claim is however less radical. He argues not that there are no differences at the neural or cognitive levels but that any such differences do not make any difference to the *reasoning ability* of delusional patients. He says that a delusion is "developed through the operation of normal cognitive processes." He continues

decades of well-controlled research have failed to demonstrate that the reasoning ability, deductive or inductive, of patients with delusions differs significantly from that of the general population. This is largely due to the fact that *by prescriptive standards*, syllogistic and Bayesian, the reasoning of the general population falls well below textbook models. Delusional subjects do not reason well and neither does the general population … the critical difference is that *the experiences of deluded patients differ from those of the general population*. (my italics) (Maher, 1999) p18).

So Maher's argument is essentially a *modus tollens*.
Either: delusion is caused by anomalous experience
Or: delusion is caused by faulty reasoning
Delusional reasoning is not faulty when compared to that of normal subjects on standard tests of reasoning competence.

Therefore: delusion is caused by anomalous experience.

Responses to the argument come in different forms. One way to reject it is to try and show that delusional reasoning *is* faulty. Still another is to show that even if reasoning *ability* is intact in delusional subjects they cannot make use of that ability or make use of it in a different way to non-delusional subjects when they try and reason about experiences implicated in delusion. This would explain why their performances on standard tests of reasoning are in the normal range for non-delusional topics. The first strategy is a competence account of the second factor, the second is a performance account.

Performance accounts. Endorsement and Explanation.

Timothy Bayne and Elisabeth Pacherie have helpfully distinguished between "endorsement" and "explanationist" views of the second factor (Bayne and Pacherie, 2004). The distinction arises in discussion of empiricist accounts of monothematic delusions such as the Capgras delusion. Empiricist accounts explain the content of the delusion in terms of the content of experience. For example the Capgras delusion is explained in terms of the experience produced by characteristic mismatch between identity and appearance. The subject "sees" the familiar but does not generate the autonomic response characteristic of recognition. This mismatch generates an error signal *explained* by the delusion.

The concept of empiricism implies that mismatch is *experienced*. In the previous chapter we gave the example of the experience of hyperfamiliarity in dreams and the Fregoli delusion as an instance of mismatch between visual and affective *experience*. However on some recent accounts the patient is not consciously aware of this mismatch. The delusion arises as a response to mismatch prior to conscious awareness (Coltheart et al., 2010).

To remain neutral I will use the term sensory processing to describe perceptual and affective processes that generate these kinds of mismatch.

The target of doxastic second factor accounts is the cognitive process that *explains* the sensory processing mismatch. We noted above that second factor theorists conceive of this process as belief fixation, which then raises an important issue addressed directly by Coltheart et. al. Belief fixation does not operate *directly* on the outputs of sensory processes but on candidate beliefs which potentially explain them. So there is an intermediate step between sensory processing and belief fixation which consists in the production of a candidate belief .

Bayne and Pacherie describe empiricism not quite in Maher's terms as the idea that explanation of experience exhausts the explanation of delusion but as the idea that the response to experience is " broadly rational." Remaining neutral on the question of whether sensory processing is experienced prior to the delusional thought we can express the essential empiricist idea this way. Delusions are basically "bottom up phenomena". Their contents preserve, or are very closely constrained by, the contents of sensory processes. In this respect they are similar to everyday empirical beliefs formed on the basis of sensory information. They are not quite " seeing is believing" because delusions do not simply report the content of experience, they amplify it slightly but in familiar commonsense ways. They are like the beliefs of someone who thinks "I have a cold" when they wake up with a high temperature and a runny nose. There is some reasoning involved here but not explicit theoretical inquiry involving generating, testing and rejecting hypotheses. Rather a thought very closely constrained by sensory information springs to mind immediately and is accepted without further inquiry. In this sense delusions remain "broadly rational" in the same way as commonsense explanation of experience.

In order to support this idea they make the distinction between beliefs which endorse and beliefs which explain experience. Beliefs which *preserve the content of experience endorse that experience*.

The basic idea of empiricist and one stage accounts like that of Maher is that delusions preserve or stay close to the content of sensory processing. Considered as abductions they are not very creative. It is in that respect that they endorse rather than explain experience. It is an additional question whether such endorsements involve problems in belief fixation. Maher argues that they do not.

The distinction between endorsement and explanations is part of a defence of the idea that endorsement is a " broadly rational" response to experience which consists in believing a default thought which preserves the content of experience. If this is correct then the difference between delusional and non-delusional must essentially be a performance matter.

And indeed this is the solution proposed by Bayne and Pacherie (2004). They say that subjects with the Capgras delusion "have the requisite inferential competence [but] are unwilling to employ it. "p66). This unwillingness could be a product of *context* or *motivation*.

Context and motivation are both of course *performance* factors and clearly interact especially given the role of emotion in both factors. The child who can unselfconsciously play *Fur Elise* with the friendly piano teacher may not be able to play as well in her music exam. Alessandro del Piero did not miss penalties for Juventus v. Bari but he could not hit a barn door in the world cup final. Such people have competence but the interaction of emotion and context affect performance. Similarly, argue Bayne and Pacherie, the affective climate in which delusion of misidentification occurs and the context (recognition of a familiar) *prevent* the subject from subjecting her delusion to "canonical procedures of verification".

In delusions of misidentification the subject sees a stranger who appears extremely similar, if not identical, to the familiar. The delusion, although it elaborates this content slightly, appears to preserve the essential structure of mismatch between appearance and

identity produced by anomalous feature binding. We can say the same thing about "hearing voices" in schizophrenia: the content of consequent delusions of thought insertion preserves the essential mismatch between the perception of self-produced actions and the absence of a signal that the action is self-initiated. In cases like this the essential idea of endorsement theories, that content reflects sensory information, seems correct. The notion of cognitive bias is invoked to explain why disconfirming evidence is not sought or assigned high probability rather than to explain the content of the delusion itself.

There are other cases in which biases are *are* plausibly invoked to explain content. Most obviously cases in which the relevant perceptual or sensory experience has an indeterminate quality. For example an intense mood, such as anxiety does not have a determinate content but nonetheless influences cognition. A highly anxious person tends to attend excessively to the possibility of threat. Anxiety-induced paranoia may be the result. A paranoid person will generate interpretations of events whose content goes well beyond anxious mood (Bentall et al., 1995, Bentall et al., 1991b, Bentall et al., 1991a, Bentall et al., 1994, Kaney and Bentall, 1989, Kaney and Bentall, 1992, Kaney et al., 1992, Kinderman et al., 1992)

The concept of performance bias is actually neutral on the question of whether or not the delusional content *preserves* the content of sensory information. Sensory information may play a very direct role as in the endorsement account of delusions of misidentification, which involve sensory information with a determinate content. Or information with less specific content, such as moods, may play a more indirect role in biasing reasoning. There may also be cases in which biases determine content *independently of sensory processing*. For example someone might have a bias to preserving sensory information *come what may*. Such a person's empirical beliefs would always endorse the content of sensory processing.

All three of these options are canvassed by Stone and Young in their classic interactionist account of delusion formation. They develop what they call an "attributional style" or

reasoning bias account which can be made consistent with both endorsement and explanationist accounts of delusion. They do not directly address the question of the role of sensory information in generating attributional style but I argue below that attributional style typically does depend on sensory processing.

Interactionism, explanationism and attributional style.

Common to both endorsement and some explanationist accounts is the idea that performance factors produce delusions. The difference is that in endorsement accounts performance factors produce a delusion consistent with the content of the sensory information, and in explanationist accounts they produce a delusion whose content exceeds sensory content. Stone and Young's account of the Cotard delusion is the best example of the explanationist strategy (Stone and Young, 1997).

Their two-factor account suggests that reasoning biases reflect two overarching principles. The first is *conservatism*, a bias in favour of preserving a pre-existing set of beliefs. A new empirical hypothesis will be assigned a low probability if it conflicts with an existing set of beliefs, tend to be reinterpreted to fit the existing conceptual scheme, subjected to a disconfirmation bias and, if it survives these obstacles, may even be compartmentalised: that is partitioned from the main theory to become an island of belief in an ocean of doxastic holism. In most situations this conservatism is adaptive. Constantly reweaving the web of belief is cognitively expensive.

Conservatism coexists with another principle of observational adequacy: a default assumption that automatic perceptual systems are producing experiences whose content faithfully represents the world. It is adaptive for beliefs to preserve the maximum amount of observational evidence (unless we are in an Ames room or flying by instruments). This leads to a general tendency to take sensory information at face value.

Normally conservatism and observational adequacy are complementary. Perception confirms rather than disconfirms our existing model of the world. "Seeing is believing", is a very sound principle for everyday life. However when observational adequacy and conservatism conflict there is no principle *internal to prescriptive theories of rationality* which can resolve the matter. Bayesian theory for example tells you how to produce the most probable set of beliefs given prior beliefs but not how to choose between two equiprobable sets. Such dilemmas are most likely to arise for extremely compelling perceptual evidence that does not fit well with existing beliefs.

Just as prescriptive theories of rationality cannot tell you which extra-rational bias to adopt it seems that the human mind has no built in algorithm for the problem. Perhaps it is an idiosyncratic matter, a product of personal and social context.

According to Stone and Young, prior to a delusional experience the probability assigned to the hypothesis, for example, that a familiar could be replaced by an imposter must be almost nil. Thus conservatism suggests that the delusional experience should be discounted, reinterpreted, rejected or ignored because it is inconsistent with the rest of the delusional subject's beliefs. Observational Adequacy suggests that sensory information should be preserved by any explanatory hypothesis. In these cases sensory information is that the identity of the seen person and her appearance do not match. This mismatch produces a default thought such as the Fregoli or Capgras delusion which preserves that mismatch. Stone and Young suggest that the delusional subject has a *bias* in favour of observational adequacy at the expense of conservatism. As a result the delusional subject accommodates the sensory information by producing the circumscribed and monothematic delusion.

A bias toward of observational adequacy fits with the endorsement account of Bayne and Pacherie. Stone and Young then need to explain why the delusional subject favours observational adequacy rather than conservatism. Here they invoke attributional style accounts of explanatory bias.

Attributional style and the Cotard delusion

Attributional style accounts emphasise the interaction between selective attention to, and recall of, particular types of experience and subsequent explanation of that experience. For example subjects with paranoid delusions focus on threat-related information and tend to attribute negative experiences to external factors rather than themselves (Garety and Freeman 1999; Bentall 2004). The opposite tendency has been observed in people with depression-related psychoses (Seligman et al., 1979). Depressive subjects focus on themselves rather than the external world and tend to attribute their distressing experiences to factors about themselves (Kinderman and Bentall, 1997, Kinderman and Bentall, 1996, McKay and Cipolotti, 2007, Gerrans, 2000).

Stone and Young's explanation of the Capgras delusion is that an intractable experience of misidentification is interpreted and explained using a paranoid attributional style. The delusional subject is " drawn toward forming an account in terms of imposters because of a more *general tendency to attribute negative events to external causes*" (p344). This idea is consistent with evidence for the presence of externalising attributional biases in subjects with delusions of misidentification, and, more generally, of attributional biases in delusional populations. This is especially so in persecutory delusions (Kaney et al., 1992, Bentall et al., 1995, Kaney and Bentall, 1989, Lyon et al., 1994). Once the imposter hypothesis is in place it is difficult to dislodge, especially since each encounter with the familiar produces only confirming evidence. Since the bias is toward observational adequacy the possibility of disconfirmation using background beliefs recedes with each instance of the experience of misidentification. A vicious circle is established. "The preferential encoding and recall of delusion- sensitive material can be assumed to continually reinforce and propagate the delusional belief." (Gilleen, 2004).

If the attributional style account is correct different attributional styles should produce different beliefs as responses to the same or similar experience. In support of this idea Stone and Young describe cases in which subjects with brain damage who show abnormalities in face processing face develop, not the Capgras *but the Cotard Delusion*.

The Cotard delusion is a delusion of negation or bodily inexistence. It is often expressed as a belief that the subject has died but a person with the Cotard delusion might claim that their body has disappeared or that they do not exist. Historically it has been associated with extreme depression and often resolves when depression recedes. Cotard, who described the delusion and noted its correlation with depression acutely noted the presence of an internalising, self accusatory attributional style " le malade s'accuse lui meme" (Cotard, 1882).

There are different interpretations of this association between experience of depression and the content of the Cotard delusion. Clearly one cannot experience being dead or being disembodied. However the experience of depression can be plausibly associated with a feeling of disembodiment (Young, 2008, Young, 2011, Gerrans, 2012). On most counts of emotion the negative affect felt in depression is awareness of an organismic state produced by the shutting down of very basic mechanisms which prime the organism for positive engagement with the world. This very fundamental mode of retreat has a bodily signature which is sensed by the patient as a "core affect". More accurately perhaps it is sensed as a lack of positive "core affect" normally present as a background against which salient experiences come to attention (Ratcliffe, Ratcliffe, 2004).

This interpretation is supported by some earlier accounts of the Cotard delusion in which patients describe progressive disembodiment. The body decomposes or fades away until the patient is " just a voice, and when that goes I won't be anything at all". In some cases the patient conceives of herself as a disembodied perspective on a lifeless cosmos, "a dead star". In these cases the patient reports, not changes in herself, but changes in the universe, one component of which is her body, an inert physical substance which first decays then disappears. Such patients will often not speak of themselves in the first person (Enoch and Trethowan, 1991).

On the basis of cases like these I once proposed a one-factor account of the Cotard delusion. In effect I was arguing, with Maher, that the experience of extreme depression explained the content of the delusion in the absence of any reasoning deficit. (Gerrans,

2002b, Gerrans, 2003, Young and De Pauw, 2003). However an endorsement account meets both these criteria while allowing that some cognitive bias is also involved. In fact one can see the depression as introducing an attributional style which prevents disconfirmation. This is consistent with most cognitive accounts of depression which treat it as a form of extreme negative appraisal bias at all levels of the supervisory hierarchy .

This endorsement account however is at odds with Stone and Young's account of cases of the Cotard delusion *following misidentification experience introduced by brain injury*. They describe cases like that of KH, a subject with brain injury who *alternated* between the Cotard and the Capgras delusion: "when he was depressed KH interpreted his altered experience as a change in him (thinking he must have died). When suspicious he interpreted his altered perception as a change in others (thinking they had been replaced by imposters)" (Young and Leafhead, 1996) p346. In this passage Stone and Young associate the internalising attributional style with depression and the externalising style with suspiciousness/paranoia.

The point to note is that the precipitating sensory information, the mismatch between appearance and identity, remains constant while the delusion fluctuates. This suggests that the content is determined almost entirely by attributional style *and* that attributional style is not dependent on sensory information or experience.

What is so striking about these alternation cases is the suddenness of change in attributional style. Normally attributional style is acquired over time. A person *becomes* paranoid or introjective and self-accusing over a protracted developmental history. Once acquired, an entrenched bias is difficult to overcome because reasoning and cognitive style are mutually reinforcing. Formal models of attributional style which treat it as a form of theory construction reinforce the point. On these accounts theories are webs of propositions which exist in probabilistic equilibrium: one new hypothesis, no matter how persuasive, should not be able to instantaneously reweave the web.

Yet Stone and Young say that KH's attributional style fluctuated with his mood.

> The key factor seems to be the patient's moods. When in a *suspicious mood* they think that others are imposters, when depressed they think they are dead. The underlying basis of both delusions could therefore be *quite similar*: a delusional interpretation of altered perception (especially *loss of affective familiarity*. (M&Lp65) my italics.

These cases of alternation between Capgas and Cotard delusion are extremely puzzling unless attributional style fluctuate with mood. But as we noted attributional style is normally explained in terms of a strongly woven pattern in the web of belief. Such patterns should not be able to be rewoven instantaneously.

Bias and default thinking.

This is less of a puzzle if the delusional belief is not treated as a theoretically anchored hypothesis but a default thought triggered by sensory processes or experience.

The reason is that imagination is *intrinsically labile*. Imaginative associations can be triggered and elaborated by experience, and then cancelled by another experience that triggers another association. Everyone is familiar with thoughts produced by moods or strong emotions, which vanish with those affective states. Children provide many examples: they are especially prone to develop trains of thought which are not based in any stable reasoning bias but triggered by current experiences.

The idea that delusions are simply default thoughts has several advantages. It would explain both their stability, in the face of repeated similar experiences, and their lability in different contexts. It would also explain the influence of particular moods, understood as core affects, which provide the background phenomenology against which particular information become salient. Moods can fluctuate against the background of a more stable attributional style, triggering uncharacteristic patterns of thoughts. This might explain

cases of alternation following brain injury as depressive mood fluctuates against a stable experiential background. Butler for example described a patient who alternated daily between Cotard and Capgras delusions (Butler, 2000). In this case it seems that the Cotard was present on waking but receded during the day to be replaced by the Capgras.

Of course this way of explaining things faces the problem of accounting for the maintenance of the delusion. The fact that a default thought originates as a response to sensory information does not explain why it is maintained. After all many thoughts are triggered by perception and can be elaborated by default processing but most are abandoned. For example, to take a case introduced by Currie and collaborators, the sight of a friend who has lost a lot of weight may prompt the thought " He has cancer". We may then continue the train of thought.

However if that thought is a genuine abductive hypothesis it will be treated as raw material for belief fixation and reconciled with other evidence and background beliefs. Unless we get some further evidence it will be rejected. More likely he has been ill or on the Atkins diet. Eventually we fix on the right explanation.

If we are to explain delusions as default thoughts we need to explain not only their origins in the default processing system but their maintenance. There are two possibilities. One, favoured by doxastic theorists, is that they are treated as abductive hypotheses but performance or competence problems with belief fixation ensure that they are not rejected but accepted.

Another possibility is that these default thoughts are not processed as abductive hypotheses at all. They retain their status as narrative elements but go on to play a role in structuring the agent's psychology.

This possibility provides an alternative interpretation of one of the strongest arguments in favour of the doxastic account: hypoactivity or lesion of dorsolateral regions associated

with delusion. This has typically been interpreted as introducing an impairment into the belief fixation process. This interpretation is a competence account of delusion formation.

Competence Accounts: deficits of belief fixation

Max Coltheart concludes his discussion of the role of right hemisphere functioning in delusion as follows :

it is specifically frontal right hemisphere damage that is the neural correlate of the impairment of belief evaluation argued here to be present in all types of delusional belief (Coltheart, 2007 p1052).

Coltheart in effect suggests that inability to activate the right prefrontal hemisphere will lead to failure to evaluate an abductive hypothesis produced as a response to sensory processing anomalies.

Coltheart's account of the role of the right prefrontal cortex as a mechanism of "belief evaluation" is consistent with the role we identified for dorsolateral circuitry as the mechanism which enables the highest levels of decontextualised supervision. One form of such decontextualised supervision is testing beliefs for consistency. When these circuits are lesioned hypothesis testing is compromised.

This also suggests that, providing default circuits are intact the psychology of a person with dorsolateral lesions will reflect the normal transmission of information through the supervisory hierarchy. Sensory processing abnormalities will be referred to the default system for supervision, triggering default thoughts. As default thoughts they are subjectively adequate responses to experience constructed as narrative elements or fragments.

If we accept that delusions start life as default thoughts produced as narrative elements then the doxastic theory needs to explain how these default thoughts become beliefs about the causes of experience.

The first stage is to treat the default thought as an abductive hypothesis. The second is to evaluate that hypothesis for empirical adequacy. A necessary condition for this process is communication between dorsolateral prefrontal circuitry and the default circuits which represent the initial default thought. When these circuits are active a patient can decontextualize. That is, treat the thought as a descriptive statement which may or may not be consistent with other beliefs and relevant evidence. The process of testing and rejecting can then proceed.

In the most drastic cases discussed by Coltheart this communication is impossible because dorsolateral circuitry is lesioned. In other cases dorsolateral circuitry may intact but mechanisms which enable its activation are suppressed or intermittently active.

A minimal description of such cases consistent with the cognitive and neural evidence is that decontextualised supervision of default thinking is absent or compromised (depending on the nature and degree of the problem in dorsolateral regions).

As a result the delusion is maintained and goes on to structure the agent's psychology. That is to produce behaviour and patterns of thinking based on the delusion.

Would a doxastic theorist be happy with this explanation, which, after all, does not mention the concept of belief or invoke reasoning deficits or biases?

Many, if not most, philosophers and some psychologists would find this minimal treatment unsatisfactory. The reason is that they would find it difficult to understand how the thought "my wife has been replaced by an imposter" could structure psychology in this way unless the agent *believes* it. On this functionalist way of thinking a representational state intermediate between perception and behaviour which structures

action and thought *just is* a belief. Beliefs are the mental states intermediate between perception and action that play a causal role in producing behaviour. Furthermore if it is a belief it must stand in some rational relationship to other beliefs and to the evidence available to the agent. Beliefs are sensitive to new evidence and background beliefs. If this is correct the delusions are beliefs and must be suitable objects for explanation by theories of belief fixation. The debt happily incurred by the doxastic theorist is the need to explain why the delusional belief is insensitive to disconfirming evidence.

For the doxastic theorist paying this debt must be seen as an investment in the project of understanding the nature of belief fixation.

I think that we should default on this debt. Rather than trying to repay it we should live within our cognitive means and content ourselves with the minimal cognitive description of the phenomenon. Namely delusions are default thoughts unsupervised by decontextualised cognitive processes. Consequently their explanation requires theories not of belief fixation but of the relationship between default processing and decontextualisation.

This minimalist view incurs a different type of debt based on the functionalist way of thinking about belief. It starts from the fact that default thinking evolved to enable stimulus independent thought. As a result it is essentially an *imaginative* process. The problem identified by the functionalist is that we do not normally think that episodes of imagination structure an agent's psychology outside of very specific contexts such as play, make believe in art, pretence and deception. Who leads a life based on imagination? The very fact that a life becomes based around a delusion suggests that the default thought has become a belief. Even if the delusion starts as an imaginatively generated default thought it is transformed into a belief as it comes to play a controlling role in the agent's psychology. Consequently delusions are beliefs.

This problem is the topic of the next chapter.

IMAGINATION INC.

The "default thought" theory of delusion needs to show how default thoughts, which are essentially simulations, can come to occupy the functional role of belief, or at least enough of the functional role of belief to cause an interpreter (perhaps even a self interpreter) to conclude that the subject is acting on the basis of belief.

The problem also arises for imaginative theorists of delusion who argue that delusions are produced by processes of imagination rather than belief fixation (Currie, 2000a, Currie and Jones, 2006, McGinn, 2004). My solution is similar in some respects to that of imaginative theorists like Tamar Gendler and Andy Egan (Gendler, 2011, Gendler, 2008b, Gendler, 2008a, Egan, 2009). Both Gendler and Egan argue that the everyday concepts of belief and imagination are inadequate for cases like delusion in which imaginatively-generated states structure psychology in a belief-like way. Gendler suggests that we need a concept of "alief" to capture these ambiguous cases and Egan proposes a concept of "bimagination".

While I think that Gendler and Egan are right about the structure of the problem their solution is to revise the personal level conceptual scheme by introducing a hybrid concept. Concepts like "alief" or "bimagination" abstract from the cognitive processes that generate the ambiguity. The approach I prefer is to trace the flow of information in the control hierarchy: to show how default system functioning in delusional mode can structure personal level psychology in characteristic ways. So while I do not disagree

with these imaginative theorists about the nature of the problem my approach is more bottom up, or "middle out" (Arbib, 2003). It starts with the cognitive level: the information processing properties of the default system. As in previous chapters on default thinking I argue that once these cognitive properties are fully described the regimentation of personal level vocabulary becomes a secondary issue. The nature of default processing explains why people are tempted to describe delusions variously as beliefs, aliefs, imaginary and bimaginary states. I think we should describe them as per the theoretical definition given earlier. As states of an unsupervised default system monopolized by hypersalient information.

The default thought theory of delusion is congenial to that of imaginary theorists like Shaun Nichols who treats imagination as the offline simulation of other mental states. This accounts treats imagination and belief fixation as black boxes in a *model* of high level cognitive processing and then employs that model to interpret personal level phenomena and explain puzzles about the relationship between belief and imagination (Nichols, 2008, Nichols, 2004, Nichols, 2006). The description of the default system provided in previous chapters suggests that the mechanism that produces simulations of personally relevant scenarios is the default system. So, for the purposes of this chapter, I will treat default thinking and imagination (as conceived by Nichols) as different levels of description of the default system involved in delusion.

With that stipulation in place I consider debates between imaginary and doxastic theorists of delusion and argue, not so much that imaginary theorists are right, but that the default thought account of delusion can explain the ambiguous properties of delusions which generate that debate.

Incorporating Imaginative States

What then is going on? The patient simply imagines that p (has the default thought that p) and accords accordingly. This type of acting *on the basis of* a mental state without metacognitive evaluation is *incorporation*.

Incorporation is in fact quite a familiar cognitive phenomenon. Any time we use the preposition "on the basis of" in psychological explanation we are referring to instances of incorporation. For example we fix beliefs, emote, form desires and act "on the basis of" perception. Similarly we may proceed from thought to thought or thought to action "on the basis of" the initial thought. In each case mental states and actions associate in a sequence.

In the case of delusion people act "on the basis of" a default simulation: an imaginative state triggered by a sensory or perceptual anomaly. Before I defend the idea in more detail consider the following case of incorporation.

A parent has a teenage child who has not returned home after a party. The thought occurs "she has had an accident" This thought is a default thought triggered by surprisal. The child is not home at 4am, which violates a prediction. There are different ways such a thought can lead to behaviour, depending on the mode in which the mind processes it. In doxastic mode the mind treats the thought as an abductive hypothesis, which, if it were true, would explain the situation. The mind then needs to evaluate it accordingly. If it is judged to be a reasonable hypothesis according to prior beliefs and the current evidence then that supposition becomes an explicit belief. Such beliefs then ground enduring multitrack dispositions to behave, to infer to emote, based on the content of the belief. And in fact this progression from imagination to belief is a natural way to reconstruct belief fixation. Bayne and Pacherie (2005) make a similar point in their discussion of what they call indicative imagination

> Indicative imagination, by contrast, is routinely triggered by

> perception. One looks at a colleague's face and imagines that they are anxious… [such]indicative imagination is continuous with belief. To imagine P in the indicative sense is to think that P might actually be true, it is to take P as a contender for the truth. Someone who leaps to conclusions—someone who fails to evaluate a hypothesis in a careful and considered manner—might easily proceed from regarding P as a contender for the truth to thinking that P is true. (Bayne and Pacherie, 2005 p170)

However it is not obligatory to treat indicative imagination this way. Someone who sees p and thinks q *on that basis* need not do so in virtue of thinking that q might be true.

To return to the case of the anxious parent. Worrying about the child, self-recrimination, telephoning the child's friends can all occur "on the basis of" the initial thought. Such associative transitions do not necessarily require the adoption of the thought as a candidate belief. This type of incorporation represents an attempt to palliate imaginatively-generated anxiety rather than fix a belief.

The point of this example is just to show that from the point of view of an interpreter incorporated imagination and belief may be very difficult to distinguish. This can also be true from the first person perspective. It may not be clear to the anxious parent herself whether she actually judges her child is in danger. And one therapeutic technique employed in cases of anxiety disorder is to gently ask the patient if she *believes* that the disaster is imminent. The typical response is that the patient cannot *not act* as if it is true or take that risk. Such answers suggest, not that the patient believes the thought, but that they cannot help acting on the basis of it. They cannot *unincorporate* it.

Belief and Imagination. Congruence and Incongruence.

It will always be difficult to establish on behavioural grounds alone whether an incorporated default thought is a belief or an episode of imagination. However there are important differences between belief and imagination that flow from the cognitive

properties of the default system, which is essentially a simulation system. In contrast systems recruited for belief fixation have cognitive properties that support the construction of true or probable hypotheses.

Beliefs stand in rational relationships to other beliefs and mental states. They are given evidential support by perceptual states for example, and have a rational relationship (*ceteris paribus*) to emotions. They confirm and disconfirm other beliefs. Beliefs are rationally constrained and their causal properties reflect this constraint. The aim of beliefs is to fit, describe or map the world. This direction of fit as it is known provides beliefs with their *congruence condition*. Beliefs are congruent when they fit their objects. This is why belief fixation is constrained by norms of consistency.

In fact all world-directed mental states have congruence conditions that prescribe goals for optimal functioning. Perception aims at veridicality or accuracy (or, if you are an enactivist successful control of action); emotions are congruent when their object has the emotionally relevant property (the congruence condition of fear is danger for example); desires have satisfaction as their congruence condition. Thus the requirement that beliefs be justified in terms of procedures which conduce to truth or probability, the correctness condition as it is sometimes called, turns out to be a special *epistemic* case of congruence.

Different cognitive processes have different computational properties that enable them to meet their congruence condition. These properties provide the *intentional structure* of representations produced by different cognitive processes. For example, the representations produced by the visual system are 3D coloured scenes derived by processing spectral and luminance information. We cannot see objects in the dark in virtue of the intentional structure of vision. The intentional structure of haptic perception allows us to feel objects in the dark.

Imagination uses the mind's cognitive resources, such as perceptual, doxastic or and emotional processing to create simulations. It thus inherits the intentional structure of these counterpart processes. However *qua* simulations imaginative states do not have

congruence conditions. This is an adaptation for planning.

In order to manage the tricky job of constructing and associating novel representations in a variety of modes, which retain their intentional structure, imagination uses cognitive processes that are stimulus-independent, not subject to norms of congruence, associative and potentially under volitional control. These conditions are necessary and jointly sufficient for a state to be an imaginative one. While they might seem obvious for standard cases (imagine being taller or wealthier: a thought which is incongruent, stimulus-independent, associative and voluntary) this definition requires some comment.

Stimulus independence is obvious for many cases of imagination (imagine lying on the beach instead of reading these words) but there are many cases where imagination is triggered by a stimulus. Often imaginary thoughts or emotions follow immediately from perception, and books and movies provide other examples of this kind of semi automatic imaginative association. Triggering of imagination however is not a counterexample to the stimulus independence condition. The thoughts in question do not *require* the presence of a stimulus to generate them. They can be generated in the absence of their characteristic object. Default thoughts in delusions of misidentification are examples. The triggering of the delusion follows immediately and automatically from the feature-binding mismatch. It thus *seems* quite stimulus-dependent. This is not however quite right. The thought "my wife has been replaced by an imposter" does not require the perception of a person. One could think that thought in an empty room for example. *Seeing* a familiar/imposter however requires the presence of the person.

Similarly the associativeness of imagination seems obvious but there are also many cases in which the relations between imaginary thoughts seem quite constrained, often by the causal nature of cognitive processes. The transition from seeing the familiar to the delusions of misidentification is an example. What makes this associative however is the contrast with the rule-governed nature of counterpart processes. For example one cannot *see* a falling object rise. One can however associate any trajectory with any object in imagination. Similarly one cannot judge according to *modus ponens* and then disbelieve

the conclusion. One can however imagine that the conclusion is false, associating a different conclusion with the premises. For a counterpart state with congruence conditions, transitions between representations are rule-governed, *even if the rules are not explicitly represented by the counterpart process.*

Similarly voluntariness does not mean that imaginative states must always be intentionally-produced. Rather it means that for the content of any state produced by a counterpart state we are free to imagine its opposite. One cannot help but see an object in the visual field according to the intentional structure of vision, but we are free to imagine it not being there or having different properties. The same is true for belief produced by judgement. Having judged that a proposition is true according to canonical norms of reasoning one cannot then judge that it is false. One cannot change beliefs unless the relevant evidence changes. But one can imagine contrary to evidence.

In fact this type of imagining contrary to current evidence is the essence of what can be called *supposition*. Problem solving often involves the generation of alternatives and testing of hypotheses that may be false or inconsistent with current evidence. We need hypotheses that go beyond the perceptual evidence, in many cases well beyond the evidence, and imagination provides them. Belief fixation tests them for consistency and accepts or rejects them. Once suppositions are judged to be correct according to the canonical norms they become beliefs we cannot then revise unless the evidence changes or we detect a mistake in reasoning. Crick and Watson were free to suppose anything about the shape of DNA in the process of generating hypotheses, but only those suppositions consistent with evidence could become judgements. Judgements are essentially responsible to epistemic norms. Imaginative supposition is not.

Imagination and its counterparts interact smoothly in a normal cognitive economy. Most famously reading, dramatic performance, pretence and play involve a smooth interplay of imaginative and counterpart processes. Planning and deliberation involve the sculpting of imagination by belief (we hold many facts constant while varying some possibilities in imagination). The smoothness of this interplay is provided by the properties of the default

system. It is specialised for stimulus-independent, associative, potentially voluntary simulation of counterpart states ungoverned by congruence conditions. Without such a faculty we would be very rigid creatures

Equally unless the representations produced by such simulations had the same intentional structure as counterpart states they would not be much help. Imagining playing the piano improves performance because it rehearses the necessary counterpart processes which can then be applied in the right context. Similarly supposition enables us to form correct conclusions because it rehearses aspects of belief fixation.

Thus simulation provides representations which can then be incorporated, that is become the basis for action (including mental action). The crucial question is whether *in incorporating a simulation one thereby believes it*. It cannot be a necessary truth that behaving *as if* a proposition is true entails that one believes that proposition.

Joint Incorporation

Incorporation is all that is needed for most of mental life. Indeed imagination and belief can be jointly incorporated provided they don't produce any drastic disruptions. Perhaps imagining I am a very good tennis player makes me more confident and relaxed when I play, which in turn makes me play better, thus producing evidence for beliefs about my standard of play. When I miss a shot I imagine it was bad luck or the brilliance of my opponent rather than ineptitude, building confidence to attempt the next one successfully. My take on my tennis ability is a fragile package of fantasy and reality: imagination and belief. Only when the coach dispassionately dissects my game do I realise how much of my tennis personality is borderline. Note however that the process of doxastic evaluation need not be counterproductive. Perhaps I have incorporated negative imaginings into my game that make me reluctant to attempt some shots and these can be replaced by positive beliefs about my ability in those areas. A necessary prelude is imagining myself performing the difficult shots successfully, leading to practicing and eventually believing that I can play them. This might be what coaches and therapists and "sports psychologists"

mean when they talk of "fragile" confidence. What they mean is that cognitive control of action is partly imaginary, partly doxastic, and that decontextualisated rational evaluation or sudden confrontation with counter evidence disrupts the balance between imagination and belief.

Given these facts about simulation it is very difficult to distinguish cases where a subject acts on the basis of a simulation from those where she acts on the basis of a counterpart state. Do I believe I am a good tennis player? Or do I imagine I am?

The place to start is from the most minimal description of the situation. Namely I am acting on the basis of a thought produced by a simulation process. Necessary for those simulations is activity in the default system. This is also the place to finish.

This suggestion runs the risk of equating empirical beliefs *only* with states produced by explicit metacognitive evaluation i.e. which are judged to fit the world by decontextualised processes of belief fixation. This would mean that a lot of human behaviour is not the result of belief, but of acting on the basis of imagination, perception, emotion, desire, default thinking and so on. I would be happy with this but there is a less radical alternative. Restrict the notion of belief to those representations that are situated in in the control hierarchy in such a way that they *can* be referred to the tribunal of judgement. Thus some default thoughts can be said tacitly believed because ultimately they can be explicitly believed. On this view an incorporated default thought ("I am a talented tennis player dogged by bad luck") could qualify as a belief if it could be successfully referred to the tribunal of judgement. If not it remains a case of incorporated imagination. Note that this suggestion does not require that such referrals produce true beliefs. We saw in the previous chapter how reasoning biases and attributional biases can ensure that the process of evaluation systematically produces false beliefs.

Thus a default thought can be a belief without being explicitly judged to be true if it is *possible* for the agent to make judgements about it.

That default thoughts successfully referred are often disallowed as beliefs ("my friend doesn't have cancer, he has just been on the Atkins diet") shows that the control hierarchy works well in standard cases and that the relevant default thought can be a belief. But in pathologies such as delusion the processes of referral and supervision are almost always out of kilter, leaving the subject no option but to incorporate the imaginary state.

The metacognitive account

The metacognitive account accepts that people, like the worried parent, can act on the basis of imagination. However the metacognitive account argues that where this is the case it is because the subject *has made a mistake* about the nature of her cognitive state. She imagines her child is in danger but believes she believes her child is in danger (Currie, 2000b, Currie and Ravenscroft, 2002).

Cleary there are occasions when this metacognitive structure applies. For example we might mistake an hallucination (which is an episode of imagination) for a percept. However the mere fact of acting on an hallucination of course does not entail the presence of such a metacognitive attitude any more than acting on the basis of vision implies that the actor believes she sees.

Of course someone with an hallucination may eventually be confronted by the fact her experience is non veridical. Perhaps will come to believe that she is having an hallucination. However it does not follow in a case like this that the true metabelief corrects a false first order belief that her experience was veridical. Rather she simply comes to believe of an imaginary state to which she initially did not take a meta-attitude, that it is an imaginary state.

Similarly with a default thought triggered by experience. In order to incorporate it the subject need not take any attitude to it at all. *Forced* to take a meta-attitude the subject might make a mistake and misidentify it as a belief. Currie and Co suggest however that

in the case of delusion the subject forms the belief about her imaginary state that it is a belief. Perhaps because, as they put it, beliefs are distinguished from imaginings by being action-guiding and not subject to the will. When the subject introspects she encounters a state which is action-guiding and feels involuntary. So she *concludes* it is a belief.

It is striking, however how few clinical accounts suggest that this captures anything like the phenomenology or psychology of delusion.

Interestingly in their criticism of Currie and Ravenscroft Bayne and Pacherie (2008) make some similar points. Their aim is to defend doxastic theories but *en route* they point out that delusions, and for that matter many other mental states, structure psychology *directly*.

Delusions and default processing.

Currie and Ravenscroft note that we often have imaginatively-generated responses to experience. Their example is the thought "he has cancer" prompted by the sight of our boss whose face has become gaunt in the interval since we last saw him. They argue that the perception of the gaunt face triggers the thought via an imaginative process. They do not add but might, that such episodes can be elaborated. One might continue to imagine the weeping widower at the graveside for example and start mentally rehearsing a funeral oration or job application. Their point is just that perception prompts these kinds of imaginative associations, which can then continue in a stream of consciousness.

Doxastic theorists would describe such imaginatively-generated thoughts as abductive hypotheses: thoughts which *if they were true would explain evidence*. Thus imagination provides the candidate beliefs, or abductive hypotheses for belief fixation. Abductive hypotheses may be accepted, rejected or revised. If rejected, imaginative processes may be required to generate alternative explanations. Perhaps my friend has been on the Atkins Diet or has had an episode of gastroenteritis.

We can imagine a person who becomes *deluded* that his friend has cancer. In such a case the thought becomes entrenched and plays a belief-like role. The patient may act on the delusion and may elaborate it, constructing stories about the origin and prognosis of the cancer. At the same time perhaps the patient may also exhibit some of the subjectivity and defensive ambivalence characteristic of some delusions. Perhaps she always insists on reinterpreting the gastroenteritis as a symptom of a more sinister problem. Perhaps she describes herself as a lone iconoclast heroically resisting the medical establishment on behalf of her friend. Perhaps she cannot click on websites providing accurate information but finds her cursor drawn irresistibly to sites devoted to personal testimonies of proxy hypochondriacs.

From the point of view of the doxastic theorist this person has a false belief based on incorrect inference about external world. What should the non-doxastic theorist say?

One thing both camps could agree on is that the delusion is a default response to perceptual information. It associates the perceptual information with a thought whose content exceeds the information given by perception.

Is that default thought an abductive hypothesis generated to explain the change in appearance? (Coltheart et al., 2010). There is no doubt that it *could* certainly play that role since default thought can be recruited as part of the process of belief fixation.

Whether in fact a default thought can be tested for consistency and coherence however depends on whether the required decontextualized processing can be activated. There is preliminary evidence that in patients with schizophrenic delusions the relevant dorsolateral circuitry is hypoactive. Furthermore that hypoactivity seems to be a consequence of hyperactivity in the anticorrelated default system.

Within the framework we are describing the explanation of the Capgras delusion in these cases goes as follows. The patient sees a familiar face. Normally that facial appearance triggers the attribution of the correct identity as a result of processing within a dedicated

face recognition system. Part of that processing includes signals initiating affective and autonomic responses. In Capgras cases this processing is disrupted due to absence of a signal initiating autonomic response (detected by SCR). Consequently the patient sees a familiar person but does not initiate the correct autonomic response (Ellis and Lewis, 2001, Young et al., 1994, Young et al., 1993).

It would be extremely unlikely that the face recognition system automatically generated the attribution of "imposter" for a mismatch between familiar appearance and lack of affective response. The output of the face recognition systems is better described by Brighetti as "identity recognition of familiar faces, associated with a lack of SCR" (Brighetti et al., 2007, Breen et al., 2000, Breen et al., 2001). It seems clear that perceptual or recognitional systems alone cannot be responsible for the content of the thought.

What then generates the attribution of imposterhood? If the description we gave above of the triggering of default thought by perception is accurate then the case fits that structure. A perceptual anomaly (in this case mismatch between appearance and affective response) is referred upwards to a higher-level control system. That system is the default system, which evolved to generate subjectively adequate responses to information referred by perceptual and sensory systems. The delusion " the familiar person has been replaced by an imposter" is a default thought which is triggered by this mismatch.

The generation of the Capgras delusion fits the same framework as the earlier example of the default thought " he has cancer" generated in response to a changed facial appearance. It is an instance of "making sense" of anomalous perceptual information by generating a default thought.

We are all familiar with *deja* and *jamais vu* experiences for places (in dreams and waking experience). And we may fleetingly think "I've been here before" or "This is a duplicate of my bedroom" in such cases. Some theorists have argued that various delusions of misidentification are instances of *deja* and *jamais vu* for faces (Malloy and Richardson,

1994, Sno, 2000, O'Connor and Moulin, 2010, Devinsky et al., 2010). Whether or not this is correct the phenomenon of default thought triggered by unusual recognitional experience is not unusual.

But normally we do not elaborate, reinforce and act on such thoughts. A very plausible explanation is that the inconsistency between this default thought and our background knowledge is detected and reconciled in favour of background knowledge. That is the role of decontextualized processing.

But delusional patients *ex hypothesi* have reduced, absent or abnormal decontextualized processing. Of course it is a strong conclusion to draw from some imaging studies of schizophrenia and lesions studies (n=9) of the delusions of misidentification which deal with small numbers that *all* delusions are unsupervised default thoughts. However there is consensus that hypoactivation in right dorsolateral areas is implicated in delusion. Coltheart encapsulates this consensus this way: “it is specifically frontal right hemisphere damage that is the neural correlate of the impairment of belief evaluation argued here to be present in all types of delusional belief. (Coltheart 2007)p. 1052)” Coltheart phrases the point in doxastic vocabulary but he might accept the more neutral substitution of “decontextualised supervision” for “ belief evaluation”.

If we accept that right frontal hypoactivity or damage leads to delusion, then we are still faced with the question of *how* it does so. The answer is not to be found in terms of the intrinsic properties of those circuits alone, but in the role those circuits play in connection with the other circuits with which they interact. The anticorrelation between the default network and the dorsolaterally-regulated networks and the type of tasks which engage the two systems suggests they are specialised for different tasks. The dorsolateral system provides decontextualised supervision for the default system. Hence when it is lesioned or hypoactive activity in the default system is unsupervised.

Delusions then arise when the default system generates and elaborates thoughts triggered by upstream perceptual and sensory processes.

The dual nature of default thoughts

If delusions are default thoughts it would explain some of their puzzling features. It would explain the subjectivity of delusion, since default thinking is intrinsically a subjective process. It generates subjectively adequate stories and fragments of stories that reconcile current information with the subject's perspective on the world. This would also explain why, as Matthew Broome says, even bizarre delusions become more interpretable the more one engages with the patient's history. However this type of engagement is more a matter of entering the lifeworld of the patient as a character in a story rather than a matter of accepting an argument (Broome, 2004).

Another aspect of delusion that could be explained by the default hypothesis is the characteristic experience of delusion. Especially in schizophrenic delusion delusional mood is usually considered by clinicians to be an essential factor but this "felt" aspect of the phenomenon is not easy to capture in the language of evidence and hypothesis.

However if the default system is "hyperactive and hyperconnected" in schizophrenic delusion this phenomenology is more explicable. One way to see this is to compare the activity of the default system in delusion, normal cognition and dreams. In normal waking cognition, as we have seen, the default network is anticorrelated with dorsolateral systems responsible for decontextualized system. In fact fluent interplay between the two systems is necessary as default thoughts are tested for validity and decontextualised conclusions incorporated in the subject's autobiography. In delusion this interplay is compromised but the functioning of the default system remains embedded in the overall functioning of the mind. It receives input from perception and sensory systems. It co-operates with memory systems and is able to activate stored schemas for responding to new information. Thus although it is an associative system the patterns of association it makes are quite constrained by the way the subject is embedded, synchronically and diachronically, in the world.

In dreams by contrast the default network is hyperactive and unsupervised by decontextualized processes (which shut down in sleep). Systems necessary for volition, reflection and intentional control of action have all been switched off. Default processing in dreaming is also characterised by absence of organising input from the sensory periphery, which leaves the default network entirely at the mercy of endogenous activation by subcortical inputs. These subcortical inputs include activity in limbic and dopaminergic systems. The latter is the mind's salience system, which evolved to make some representations highly salient, to pop out of the cognitive background and command attention and engage motivational systems. Thus in dreams the default system churns through representations of scenes which often feel significant and emotionally charged. The dreamer traverses an eerie dreamscape in which she has experiences but does not seek to explain or respond to those experiences (Desseilles et al., Dawson and Conduit, 2011, Hobson, 1999, Roehrenbach and Landis, 1995).

Thus dreams represent an extreme case of hyperassociative default processing "a unique and more fully developed form of mind wandering, and therefore … the quintessential cognitive simulation." (Domhoff, 2011 p1172) The connection with delusion noted on phenomenological grounds by many patients and clinicians has been made in cognitive terms by theorists we considered in chapter 4.

This concentration on the cognitive properties of dreams also diagnoses the source of a debate about the phenomenology of dreams which parallels the debate about whether delusions are beliefs: namely whether dreams are states of imagination or belief (McGinn, 2004, Gendler, 2011, Sutton, 2004, Ichikawa, 2009, Ichikawa, 2008). My sympathies are with imaginary theorists in this debate. In dreams the control hierarchy is not configured to test representations for veridicality. Dreams have no congruence conditions. However the pathway to this conclusion is via the cognitive theory of default processing rather than reflection on the phenomenology of dreams and the conceptual contraints on personal level concepts of belief and imagination.

Thus the similarity between dream experience and delusion can be explained by the fact that they share important properties of default thinking, although delusions occur in a waking state. Namely they involve the activation of components of a network which simulates and a stream of personally relevant experiences and thoughts in the absence of actual perceptual input or supervision by decontextualised thought.

This may also account for some of the ambivalence of the delusional attitude expressed by some schizophrenics and Alexander's Capgras patient. Involuntary dreamlike experiences or patterns of thought occurring in a waking/aware state felt to be very significant but at odds with background knowledge must feel very strange. At the same time the patient (*ex hypothesi*) cannot submit them to decontextualized evaluation.

The degree and nature of the compromised interaction between default and decontextualised processing must differ from delusion to delusion according to the precise causal aetiology. In schizophrenia for example the default system is hyperactive and hyperassociative. Decontextualised systems are intact but hard to activate. In the neuropsychological delusions like Capgras, described above, where decontextualized systems are lesioned the bias to default thinking does not arise from hyperactivity in the default system. Consequently the delusion may not be very elaborate or florid (even if it is distressing). These delusions are often described as circumscribed and monothematic. Perhaps they are default thoughts intractably triggered by anomalous upstream processes which do not develop into elaborate hypersalient narratives because the default system is intrinsically in a relatively normal state but unable to be supervised by decontextualised processing. This speculation suggests that these delusions are almost the converse of ventromedial damage. Ventromedial patients can draw decontextualised conclusions but cannot take the autobiographical perspective on information required to link them to planning decision and action. Capgras patients with dorsolateral lesions are stuck with the intractable default thought and its prompt to action but cannot decontextualize, or if they can, cannot produce an autobiographical narrative which incorporates the non delusional alternative. They cannot "make it theirs" so to speak.

Imaginative resistance and the essential indexical

Hume described inability to imagine things inconsistent with self-representation in the Essay "Of the Standard of Taste" (1757), "*I* cannot, nor is it proper that *I* should, enter into such sentiments." (My italics). Hume was very clear that he was not talking about belief but imaginative identification with repugnant attitudes or acts. These are cases in which as Nichols (2004) puts it " the imagination rebels against certain elements". Another way to put this is to say that we cannot construct simulations for acts or thoughts which are too inconsistent with the fundamental cognitive and affective states which motivate us. This phenomenon has been called "imaginative resistance" by philosophers (Driver, 2008, Gendler, 2000, Levy, 2005, Todd, 2009, Walton, 2006). It seems that there are *aspects of the self which cannot be imagined away*. We can imagine some propositions being true *de dicto* (as a set of properties) or *de re* (a particular individual instantiating the properties*)* but not *de se.* We cannot imagine ourselves instantiating the relevant properties. In classic cases of imaginative resistance the relevant properties are morally repugnant (such as torturing a child).

I think that a similar thing is going on in many cases of ambivalence towards delusional content. Inability to "own" the clinical/causal explanation of their delusion can be explained in terms of "imaginative resistance. That is to say the subject cannot *imagine* that *she is* deluded. She can entertain the proposition theoretically but when she tries to make essentially indexical " the imagination rebels".

The delusional patient who is presented with the non-delusional explanation of her experience needs to be able to make it *essentially indexical* for her and incorporate it. However propositions which can be entertained in theoretical mode ("I am mentally ill") cannot be incorporated as part of a narrative the subject experiences as her own. The delusional story however is one which (even if it is disturbing in some cases) is easier to incorporate.

The doxastic theorist might reply that the delusion is easier to incorporate because it is believed. However the easiest belief to incorporate, according to norms of belief fixation, is the true one: that he patient is having a psychiatric episode. After all it fits better with all the evidence available to the patient than the delusion.

However the term "easier" should be interpreted here in cognitive processing not epistemic terms. It is not a matter of which hypothesis best fits the evidence with least readjustment of related beliefs but which pattern of incorporation makes the least cognitive demands *given the structure of the control hierarchy in delusional mode*. That mode is characterised by inaccessibility of decontextualized processes, highly salient default thoughts triggered by automatic systems whose processing cannot be easily overridden from the top down, and a narrative that accommodates those thoughts. Against this background the non-delusional thought often encounters serious imaginative resistance and becomes impossible to incorporate.

Robert Nozick once argued that a woman presented with irrefutable evidence that her son had committed a brutal murder might rationally disbelieve that evidence and cleave to a belief in her son's innocence. He meant that this reaction was plausible, psychologically explicable and forgivable. But surely the most fundamental reason for her maintaining her son's innocence is that she cannot imagine (in the strong sense of making that thought essentially indexical for herself) his guilt. Is she deluded?

Cognitive therapy for doxastic theorists

This idea suggests an alternative interpretation of cases in which patients abandon their delusion following confrontation with counter evidence and rational evaluation. Doesn't that show that delusions are beliefs? After all imagination is not sensitive to this kind of rebuttal. Some of the things delusional patients say seem to bear this out.

John Nash's description of his battle with schizophrenia is an example.

> Initially I did not hear any voices. Some years went by before I heard voices and -- I became first disturbed in 1959, and I didn't hear voices until the summer of 1964 I think, but then after that, I heard voices, and then I began arguing with the concept of the voices.
>
> And ultimately I began rejecting them and deciding not to listen, and, of course, my son has been hearing voices, and if he can progress to the state of rejecting them, he can maybe come out of his mental illness.
>
> The consequence of rejecting the voices is ultimately not hearing the voices. You're really talking to yourself is what the voices are, but it's also parallel to a dream. In a dream it's typical not to be rational.
>
> So in rejecting some of the political ideas, that had a relation to the voices, so I could think of a voice maybe as presenting what was analogous to a political argument, and then I could say, I don't want to listen to that. John Nash (my emphasis).

This suggests that Nash treated his delusion as an hypothesis which he disconfirmed.

It is interesting however that Nash treated the voices as antagonists in a political argument rather than as participants in a logical or mathematical debate. Their initial power to monopolise his attention does not seem to be a matter of proof or evidence. He once said “the ideas I had about supernatural beings came to me the same way my mathematical ideas did. So I took them seriously”.

It is important to understand that Nash is not saying that his ideas about supernatural beings arrived after months or years of intense proving and checking theorems, defending them against attempted refutation. Rather they occurred in his mind with the same intensity and spontaneity as mathematical insights. In the mathematical case those ideas survived and were strengthened by the process of proof but it is not clear that he ever

tried to prove or disprove his delusions. Rather while deluded "he took them seriously" constructing elaborate narratives of persecution in which *he* was the focus of attention by global superpowers the church the CIA and sometimes intergalactic forces. He constructed autobiographical episodes to fit the experience. In other words he incorporated his delusions.

When Nash treated his voices as a political argument it was so that he "could say, *I don't want listen to that*". He didn't attempt to disconfirm them but rather he starved them of attention. He managed to reduce their salience and created room for other default thoughts previously crowded out by the delusion.

Robert Chapmans's personal account of recovery from a classical case of schizophrenic delusion is equally interesting. When deluded be believed he was on the verge of worldwide fame as a modern DaVinci or the victim of global conspiracies to steal his ideas and deny him recognition (Chapman and Chapman, 1988).

He describes the origin of his delusion as involving preoccupations, racing thoughts, a sense of grandiosity and significance followed by intense suspiciousness. He fitted these experiences into spectacular narratives of persecution in which he was the central focus of others intentional activity. As he put it he asked "why" questions which demanded that his experience be made intelligible in terms of agency rather than "how" questions which demand specification of causal mechanisms and logical consistency.

Both Nash's and Chapman's delusions are consistent with the idea that they originated as default thoughts no doubt triggered by very vivid experiences intensified by the salience system. In both cases the delusions were incorporated and elaborated into a paranoid narrative which became quite fixed.

Against this interpretation we can consider the Chapman's explanation of his recovery, which he explains as a case of the falsification of an empirical theory. The transition, as he describes it, was a result of applying principles of cognitive behavioural therapy.

Challenging the delusional belief, gathering disconfirming evidence, developing alternative explanations and finally “ replacing delusions with the objective truth”.

Rather than imagining being spied on and following the train of associated thoughts, Chapman found perceptual evidence that he wasn’t. “ I look into walls floor radiators, ceiling ducts, and other orifices where I suspect a spying device is hidden, but I never find any” “Did I actually see it or just a “ sign” of it? Did I really hear it or could I have misinterpreted it? Did I smell it, taste it or feel it?” Chapman asked, over and over.

Chapman’s case is especially interesting because initially it seems to support the idea that the second factor in delusion is a reasoning impairment leading to a faulty explanation of experience. When he subjected his delusions to rational evaluation they disappeared.

However the fact that a delusion remitted in the face of hypothetical reasoning does not show that it *originated* as the result of an failure of hypothetical reasoning. Nor does it follow that in fact he constructed a new, better, theory to explain experience. Rather his new scientific attitude had the effect of providing new set of experiences to incorporate into his autobiography. Not only that but the reallocation of cognitive resources reduced the salience of the experiences which prompted the delusions. They receded into the metacognitive background as he focused his cognitive efforts on collecting mundane veridical experiences of the external world.

The scientific attitude was required to unweave the threads of this narrative and to make available a new set of empirical materials from which to reweave a story. Out of these materials Chapman reconstructed a story in which he was not a persecuted genius but an intrepid explorer on quest for knowledge “ I will not allow myself to be misled and deceived anymore. As much as these strange beliefs seem awfully real I will continue to investigate them.” Instead of private fears and paranoid interpretations he concentrated on publicly available phenomena: not the fear of being observed by hidden spies but the sight of the empty broom cupboard.

Clearly Chapman's hard won scientific attitude to experience was crucial to his recovery, but it does not follow that the *cause* of his delusion was a failure to take a scientific attitude to his experience. The cause of his delusion was the warping of his autobiographical narrative by hypersalience of paranoid and grandiose thoughts.

When Chapman became a robust empiricist he also became a more reliable narrator because the raw material of his autobiographical narrative changed from subjective adequate interpretations of experience to publicly available information. His story about his experiences now overlapped with that of others because it incorporated more of the same information.

Imagination and psychological structure

Imagination is a distinctive cognitive process, with its own architecture, which exploits other cognitive systems to generate representations with distinctive intentional properties. This is not an accidental feature of imagination: specialised neural circuitry has evolved to enable humans to construct and manipulate representations which have representational contents but no congruence conditions, conferring distinct advantages on the human species.

Although they lack congruence conditions imaginative episodes can be *incorporated* into an agent's psychology, structuring behaviour and patterns of thought. In these cases imagination behaves in a very similar way to belief. Incorporation turns out to be a very fundamental mental operation in which psychology is structured "on the basis" of a mental representation without it being the case that a subject takes an attitude, or even has the disposition to have an attitude towards, that representation. If the concept of incorporation is incorporated into our explanatory repertoire it becomes easy to see how imagination can structure an agent's psychology without it being the case that the agent believes the imagined content or makes a mistake about the process that produces it.

THE SENSE OF AGENCY, LOST AND FOUND. EXPERIENCE AND THOUGHT IN SCHIZOPHRENIC DELUSIONS

A book on delusions would be incomplete without a discussion of schizophrenic delusions, the most common, clinically significant, and, in some ways, most baffling forms of delusion (How can a person genuinely experience episodes of thinking that originate in someone else's mind?). Previous chapters touched on aspects of schizophrenia in discussions of neurobiological eliminativism, the default system, doxastic and narrative theories of delusion and dopaminergic mechanisms of salience. However they did not attempt to provide a systematic explanation of schizophrenic delusion. That is the aim of this chapter, which focuses on delusions of alien control and the (I will argue, related) phenomenon of delusions of thought insertion.

The standard doxastic explanation of these delusions is that they are beliefs generated to explain the loss of "sense of agency", a subtle form of bodily awareness that arises in the process of controlling action. When this sense of agency is absent the subject performs an action but does not experience it as self-initiated. This can be true even in cases where she explicitly intends the action. For example in one experiment subjects asked to move a joystick did so, which indicates that they explicitly intended the movement, but reported vivid sensations of not being the agent of the action (Mlakar et al., 1994, Spence et al., 1997a).

Delusions of control are correlated with abnormally high levels of activation in the right inferior parietal cortex (Behrendt, 2004, Blakemore et al., 2000, Maruff and Currie, in pres, Spence et al., 1997b, Ganesan et al., 2005, Langland-Hassan, 2008, Spence, 2002, Spence et al., 1997a). Spence has observed ruefully that recent high tech studies establishing this correlation merely confirm the prediction made by Angyal in 1936 that the inferior parietal lobe would be implicated in passivity experience in schizophrenia (Angyal, 1936).

The initial aim of this chapter is to explain this correlation using a predictive coding *model* of motor control. The ultimate aim is to transform that model to a neurocognitive theory by considering evidence that the neural circuitry implicated in delusions of control actually implements a version of that model. The essential idea is that sensorimotor systems "predict" sensory consequences of movements initiated by the organism. This prediction allows an organism to correct errors by comparing predicted to actual sensory inputs. Crucial dimensions of comparison for action control are between predicted consequences of motor commands, proprioceptive (the sense of bodily location) and visual feedback (Blakemore et al., 2002, Hartmann et al., 1984, Synofzik et al., 2009).

The predictive coding framework generates a model cognitive architecture for motor control, which can be represented in the diagram below (the forward model as it is known). The basic idea is that a goal-directed system computes the difference between the goal state and its current state (the inverse model). This inverse model generates an intention to produce a movement that gets the system from the current state to the goal state. That intention generates, via a downward cascade of cognitive processes, a motor instruction whose execution realises the intention. Execution of the motor instruction produces feedback in the form of sensory inputs. That feedback is "compared" to a "prediction" of the sensory consequences of the motor instruction and any discrepancy generates an error signal, which is then used to generate a new set of motor commands to correct the movement. And the process iterates until the error is cancelled. Importantly, the system is hierarchical with each level in the hierarchy using the predictive coding strategy. Error signals and corrections propagate up and down the hierarchy. My intention

to park the car for example is ultimately realised at low levels by minute adjustments of pressure on the steering wheel and accelerator (Pacherie, 2000). Failures at low levels which cannot be automatically corrected at those levels produce error signals which propagate up the hierarchy.

Different layers in the control hierarchy exploit predictive information in different ways. Automatic sensorimotor loops couple the organism to the environment by comparing actual to predicted proprioceptive feedback consequent on motor instructions. If the movement is unsuccessful the mismatch creates an error signal that is sustained until the movement is corrected.

This process of comparison is very rapid and typically stays tacit. For example in a successful grasping movement the proprioceptive representation of the actual movement trajectory is almost instantaneously matched to a prediction. Typically these corrections occur instantaneously and the relevant signals are never referred to higher levels of supervision.

The sense of agency arises in the interaction between this essentially proprioceptive feedback loop with another feedback loop used for higher level *explicit* visually-guided control or (sometimes) mental rehearsal of action. In this case explicit visual information is used to control movement and reduce the error signal. The need for this type of control arises when an error signal is referred up the hierarchy from automatic to controlled processing. This type of careful, attentionally-guided, visuomotor control requires activity in a circuit comprising at least the SMA, parietal (especially inferior parietal) cortex and cerebellum (Blakemore et al., 2002, Danckert et al., 2004, Lafargue and Sirigu, 2006) and its properties are captured by the predictive coding architecture.

When predictions are sustained in this process of high-level control we experience the "sense of agency". When higher levels of control lose access to predictions the sense of

agency is also lost. In this state the subject sees an action performed by her own body but does not experience herself as the initiator of that action. It is experienced as *passive*. This passivity experience is not like experience of involuntary movement such as a tic or twitch because movement it accompanies is an action such as a reaching or grasping. The movement is intentional but not *felt* to be intentional. The subject feels as though she is passively performing an action she did not initiate.

This minimal *passivity* interpretation of the phenomenon contrasts with richer *externality* interpretations (Bayne and Pacherie, 2007, Pacherie et al., 2006). On these interpretations the subject actually experiences being moved by another agent. This interpretation is very congenial to endorsement accounts of delusion that argue that delusions inherit the content of experience. I consider one version of the predictive coding theory which argues that the content of the relevant experience is of externality/being moved by another agent (Georgieff and Jeannerod, 1998, Jeannerod, 2006, Gallese et al., 2004, Iacoboni et al., 2005, Gallagher, 2004). While this would simplify the explanation of the content of delusion I ultimately reject it for the cognitively more parsimonious view that the experience is of passivity.

A problem arises when we extend the structure of this explanation to explain delusions of thought insertion. The reason is that if the predictive coding account is to explain delusions of thought insertion we need to reconceptualise thinking "as a motor process" as John Campbell put it (Campbell, 1999). This problem becomes tractable if we think of the thoughts in question as linguistic, as "inner speech" (Langland-Hassan, 2008, Shergill et al., 2003). The reason is that speech *is* a motor process. This option has been rejected as inadequate to the full range of thoughts and emotions experienced as alien by schizophrenics. However the more modest target of this chapter to explain a loss of the sense of agency for inner speech.

The rest of the chapter proceeds as follows. I first discuss the way visual and bodily information are integrated in motor control and note that the human mind seems to be wired to give visual information precedence in resolving conflicting sources of

information. This is no doubt a consequence of the fact that humans rely heavily on visual channels for higher levels of cognitive control.

I then describe the evidence that that architecture is implemented by distributed circuitry centred on the parietal cortex. This attempt to show that a model is actually implemented is part of the attempt to transform it from a model to a neurocognitive theory. I then argue that the passivity experience in these delusions can be explained in terms of activity in predictive coding mechanisms. Finally I consider and reject alternative explanations that treat the delusion as a defensive response to ego-dystonic thoughts.

The sense of agency. Lost and found

The "sense of agency" for actions arises at the border between controlled and automatic processing. Automatons do not have a sense of agency and, in so far as we are automatons, neither do we. The sense of agency arises in the process of transferring control of action from automatic to controlled processing. Equally, when the control of action is successfully delegated to automatic systems the sense of agency disappears. For example when descending a staircase we are not aware of a sense of agency. Our motor control system rapidly and automatically constructs a representation (forward model as it is called in the figure above) that predicts the sensory consequences of movement. In a regular descent the sensory consequences (proprioceptive and haptic) of each step are predicted by the model and hence no error signal is generated. If however a step is irregular the prediction is violated. If the error is minor (perhaps the step is a couple of millimetres lower than predicted) the proprioceptive mismatch is automatically and instantaneously corrected. If, however, this form of control is inadequate (perhaps a step is missing, or the staircase has crumbled) and our foot does not encounter the step we need to shift to higher levels of control, carefully descending by looking at each step. In this case we are using visual information to cancel the error signal generated by the proprioceptive mismatch. The sense of agency arises at this transition between controlled and automatic processing. Because this interplay between automatic and controlled

processing is constantly fluctuating the sense of agency waxes and wanes according to the degree and fluency of higher-level control.

This is why, as Pacherie puts it

> Our ordinary experience of agency may be simply a rather diffuse sense of a coherent ongoing flow of anticipations and sensory feedback. It may be that it is only when we are voluntarily attending to what we are doing, or in the case of a serious mismatch between predicted and actual outcome that we become explicitly aware of what we are attempting to do as opposed to what we are actually doing. (Pacherie, 2001)

Slightly paradoxically it seems that the sense of agency is more likely to arise in situations where the control of action is precarious, actually or potentially unsuccessful. However this is not a paradox, we *are* more aware of our agency when learning a musical instrument, walking through a minefield or threading a needle, than when performing a task automatically. This phenomenon explains the phenomenology of expertise. It is often said that the actions of highly skilled practitioners, like brain surgeons, musicians or tennis players are not consciously controlled. Perhaps a better way to put this intuition is to say that, because their actions are successful they, unlike novices, do not need to constantly rely on careful integration of explicitly represented information to deal with the consequent error signals. The novice pianist constantly experiences unmet predictions (proprioceptive, haptic and auditory) and is therefore conscious of a sensation of effort or failure. She has to resort to explicit control of her movements, carefully matching action and consequent perception to predictions generated by her intentions. The expert simply experiences only the “flow” of smooth matches between prediction and reafference managed by automatic action components.

The degree of conscious control, and hence sense of agency required fluctuates according to the difficulty of novelty of the task, level of expertise and stability of the environment. Furthermore an action, such as parking a car, has multiple components that require

different levels of attentional control. Carefully steering between obstacles at a steady speed may generate a sense of agency for steering while simultaneous control of the accelerator, being delegated to automatic processing, does not. In other conditions, such as trying to reverse without hitting a wall, the sense of agency might be produced by trying to control the accelerator while steering becomes automatic and hence phenomenologically silent.

The priority of visual experience

Not only is the sense of agency hard to isolate phenomenologically, but when there is a conflict between the sense of agency and other information, such as visual or bodily, the sense of agency is liable to be trumped by other phenomenology. This phenomenological priority of visual information over more subtle forms of bodily awareness seems to be an architectural feature of human cognition. For example in the rubber hand illusion a subject wears goggles which project an image of a rubber hand a metre ahead of the subject. An experimenter simultaneously strokes the hand of the subject and the rubber hand. Thus the subject feels her own hand being stroked but sees the rubber hand being stroked. After a short period the subject "feels" the rubber hand being stroked. It is as though the bodily phenomenology "jumps" into the rubber hand. When the experimenter stops stroking the subject's hand the subject continues to "feel" the rubber hand being stroked.

.

The experiment creates three contending sources of information: bodily/haptic, visual and propositional. The explanation of the illusion is that the supervisory representation being used to reconcile inconsistency between these sources has a strong predictive bias or "prior" which gives visual information precedence (Blanke and Metzinger, 2009, Lenggenhager et al., 2007, Schwabe and Blanke, 2007). In this condition inconsistency between visual and bodily information is resolved in favour of visual, producing the illusory phenomenology. Higher-level background propositional knowledge has no effect on this phenomenology. Versions of this illusion have been demonstrated for dolls, dummies, other bodies and inanimate objects such as cardboard boxes. In all cases it seems as if, once the expectation of consistency between visual and bodily information is

established at low levels in the control hierarchy, inconsistency is resolved in favour of the visual, despite contrary background knowledge. Thus the propositional information that it is a rubber hand being stroked has no influence on the phenomenology.

The pairing in the rubber hand illusion is between passive haptic and visual information but the mind also performs a similar matching process between predicted and actual information about bodily *movements*. An experiment conducted by Ramachandran and Rogers on a patient with phantom limb pain supports these ideas (Ramachandran and Hirstein, 1998, Ramachandran and Rogers-Ramachandran, 1996). Their patient had intractable muscular pain in the phantom limb. Their solution to the patient's problem required the patient to experience the sensation of moving the limb on the basis that movement relieves cramp. Of course this seems impossible in a phantom limb: there is no limb to move. However Ramachandran and Rogers overcame this problem using a box, open at the top, with mirrored walls. The patient was asked to place his intact hand in the box, through a hole in the side and place the forearm of his amputated limb parallel to his intact one, as though he was inserting that hand into the box as well. The patient then looked into the box from above. As a result of the way the mirrors were placed, the patient had the illusion of two hands inserted into the box. When asked to try and move the two hands in concert the patient "saw" two hands moving and his phantom muscle spasms were relieved.

The explanation is that the patient experienced a match between what he saw and what his mind predicted: namely an arm moving in the expected trajectory. A case of illusory visual information, which matched a prediction, produced by the artificial provision of visual feedback.

Rubber hand and phantom limb cases introduce the idea of the priority of predicted visual information in situations of close visual monitoring of bodily movement and sensation. However in the rubber hand illusion there is no competing source of information: because the subject lacks a limb there is no *proprioceptive* feedback to compare to visual input. The patient experiences a perfect match between visual input and prediction.

In cases of passivity experience there is an intact limb, and hence source of proprioceptive feedback but the patient seems to lose access to the relevant predictions, so no error signal propagates to higher levels of control. The patient has a conflict between proprioceptive and visual information resolved in favour of visual. At the same time the action is experienced as passive because of the loss of access to predicted proprioceptive feedback.

Predictive mechanisms and cognitive architecture

The brain has specialised parietal circuits for the integration and reconciliation of information from different sensory channels: visual, proprioceptive, motor and haptic. Integration requires the parietal cortex to encode information about the trajectory and sensory consequences of movements in a format which allows reconciliation of visual and proprioceptive information (Synofzik et al., 2008). This encoding is the "prediction" of the consequences of a motor instruction. The nature of this predictive coding and matching process is a topic of controversy but one fact seems clear. Anomalies are associated increased levels of activation in right inferior parietal cortex (Lafargue and Franck, 2009, Ganesan et al., 2005, Danckert et al., 2004, Eidelberg and Galaburda, 1984, MacDonald and Paus, 2003, Rizzolatti et al., 2006, Rushworth and Taylor, 2006, Sirigu et al., 2003).

There is an important point to note here. Activity in right and left inferior parietal cortex is *anticorrelated* in action monitoring. High levels of activation in one region are correlated with lower levels in the other. Thus the high level of right inferior parietal activation associated with passivity correlates with low levels of left interior parietal activation and *vice versa*. Since high levels of activation in right parietal cortex are associated with passivity, this suggests that activity in the left parietal cortex would be associated with *presence* of the sense of agency (Lafargue and Franck, 2009). And indeed patients with lesions in this area performed poorly on motor control tasks which required them to determine whether an observed action was one they initiated (Sirigu et al., 2003).

Consistent with this finding, disruption of parietal activation by TMS also disrupts the sense of agency (Rushworth and Taylor, 2006, MacDonald and Paus, 2003).

This suggests a mechanism for prediction. Inhibition or attenuation of activity in right parietal areas consequent on signals from the motor cortex that produce the movement. As a result incoming signals do not produce as much activation as they would if the right parietal regions had not been pre-attenuated. The experience of loss of agency arises when, due to failure to pre-attenuate parietal areas, intended bodily movements produce the same levels of right parietal activity as unintended. In this condition the unattenuated parietal activity is experienced as a movement the agent did not intend.

Because of the anticorrelation between left and right parietal areas it is tempting to localise the sense of agency to activity in the left parietal cortex. Some authors do write as if this is the case, treating left and right parietal cortex as specialised for representations of "self" and "other" for example (Grézes and Decety, 2001). However, from the point of view of view of the motor control system, left and right parietal cortex are correlated with self and other-initiated movements, not because they intrinsically represent self and other but because their response to sensory representations of movement trajectories is different according to whether those trajectories are predicted consequences of upstream motor signals. Of course when action is self initiated a motor signal will attenuate right parietal activity accordingly. When that attenuation fails the sense of agency is lost.

The sense of agency is hard to test directly. It arises as a by-product of the high level control of action. Consequently it is investigated in a range of conditions that require the implicit or explicit attribution of a movement to an agent in the context of action control. For example in experiments which ask the agent to say whether or not a cursor moving on a screen is moved by the agent, experimenter or the computer. In one ingenious experiment hypnotised patient's arms were attached to a pulley and patients were asked whether they or the pulley moved their arm.

One feature of these experiments is that they provide a context for any experience of loss of agency. The actual signal available to higher levels of control in the form of unpredicted experience is that the observed movement is not a consequence of an upstream motor signal. The subject who becomes aware of that signal *in the context of higher-level control* generates the conceptual interpretation or attribution of agency. In these experimental setups the agent who feels something amiss has a choice between self or other attributions as a response to that experience. The context creates a default attribution for the agent.

However in the normal phenomenological flux of life the context may not provide an obvious default attribution for passivity experience.

Experimental Evidence

Even when sustained, the sense of agency is not normally salient amidst the stream of consciousness because it tends to be crowded out by other phenomenology. For example the phenomenology of controlled grasping produces haptic sensations on the fingertips, the sensation of muscular contraction, proprioceptive sensation and the visual information. In amongst this complex phenomenology, competing for prefrontal resources, is the subtle and evanescent sense of agency.

Fortunately, for phenomenologists if not patients, there are always patients whose impairments help decompose the phenomenological and cognitive structure of action and isolate the relevant mechanisms. Deafferented patients provide an important resource. Marc Jeannerod describes a case of a paralysed patient trying to lift a limb (Jeannerod, 2006). The motor instruction generates a prediction of the reafferent sensory consequences (contraction of muscles, proprioceptive changes in sense of position and perception of movement) but because the limb is deafferented no such feedback is produced. Nonetheless the patient experiences a “sense of effort”.

Jeannerod's explanation is that the patient is experiencing sustained activation in neural circuits which *precede* the execution of motor instructions. Jeannerod describes these tacit processes as *covert* actions, and the action itself as overt action: the outcome, or translation into movement, of the covert processes that precede movement. Normally these covert processes are unavailable to consciousness because they are translated instantaneously to overt action and matched to sensory feedback. It is only when overt actions are unavailable, frustrated or unsuccessful, that we become aware of the covert, normally tacit, aspect of movement.

Cases of paralysis provide evidence for the idea that the "sense of effort" is a way of becoming aware, not of bodily feedback (because the patient is deafferented) but of what Jeannerod has called motor preparation or covert action. Deafferented patients can accurately sense the differences in effort involved in trying to make different actions for example and the same phenomena can be produced in healthy patients who are temporarily deafferented. Lafargue and collaborators have conduced an elegant set of experiments with hemiparetic patients asking them to compare or estimate the efforts involved in making and attempting actions with paralysed and non paralysed limbs. In the absence of feedback from the deafferented limbs patients seem to rely on the sense of effort generated by "covert" processes (Lafargue et al., 2008, Lafargue and Franck, 2009, Lafargue and Sirigu, 2006). Similarly, a patient with no proprioceptive feedback has the capacity to become aware of her efforts to produce actions (Lafargue et al., 2003). Together all these cases suggest that sense of agency is a form of awareness of the preparatory processing involved in movement rather than feedback from the movement itself or of a match between feedback and prediction. Imaging studies show that this awareness of covert processes is associated with activation in the right inferior parietal cortex.

Another way to demonstrate that there is a phenomenology of agency independent of bodily feedback is to focus on cases of *imagined* action (Decety and Lindgren, 1991, Jeannerod and Frak, 1999) . In these cases motor output is suppressed and there is no sensory reafference. As Jeannerod would put it the action is entirely covert. Yet there is a

sense of agency. Jeannerod describes awareness of agency in cases such as this as becoming conscious of covert actions, which are normally inaccessible to consciousness, as a result of sustaining activation in circuitry which normally decays almost instantaneously. The important point is that these cases dramatise the fact that covert activation decays when transformed into overt because overt action produces reafferent sensory feedback which cancels covert activity. Thus awareness of covert action is in fact awareness of this activity which becomes unusually sustained in cases where feedback is not anticipated or does not arrive.

> if motor preparation, which normally lasts for a very brief period of time could be *prolonged* the intention to act would become progressively a *motor image of the same action*. If this were the case, then *the non conscious to conscious transition* would be only determined by the time allowed for the preparation process to access awareness and to become conscious. *Actions which fail or which are cancelled at the last moment may be situations where a non-conscious program is transformed into a conscious image*.

In fact the transition is not so much from conscious to non-conscious, because covert action is not *intrinsically* unconscious. Rather the feeling of controlling an action arises when activity, which normally decays quickly, is maintained because there is no reafference to cancel it. This activation then becomes in effect an error signal which attracts prefrontal resources (Saoud et al., 2000, Danckert et al., 2002).

The notion of covert actions introduced by Jeannerod, understood as *everything* which precedes actual movement includes premotor processing, motor intentions and commands *and* the prediction of the reafferent consequences of motor processing.

Is there any way to parse contributions of different cognitive components of the covert action sequence to determine if any is especially implicated in the sense of agency? In fact there is. The neural basis of motor control is in fact quite well mapped and the contribution of different circuits comprising the descending hierarchy from intention to

movement and monitoring via sensory feedback is established by a combination of lesion and imaging studies.

The consensus is that the supplementary motor area (SMA) signals parietal areas according to the motor instruction it is executing. “A corollary of the motor signal involved in intended effort, in the SMA, could be sent to attenuate parietal activity” (Lafargue and Franck, 2009). This is called the corollary discharge or efferent copy in forward models: it “tells” the parietal cortex that a motor instruction to produce a certain movement has been issued and that consequent sensory feedback generated by the movement is reafferent rather than afferent. That signal attenuates activity in the parietal cortex so that when predicted (reafferent) feedback arrives it does not produce the same level of activation as afferent signals.

This is what is meant by “cancellation” of predictions. When inputs to parietal circuitry are predicted, activation, already down regulated, rises less than when inputs are unpredicted. When they are not predicted activation is sustained beyond thresholds sufficient to attract attentional resources which amplify it and make it available to central processes. It is this sustained activation which is sensed as the “sense of agency”.

This is why Jeannerod suggests that deafferented patients experience a sense of agency for attempted actions. Activity in the inferior parietal cortex is initially attenuated by the SMA as the motor instruction is issued but it increases and is sustained because there is no movement to produce bodily feedback which cancels activity in the inferior parietal cortex.

Thus the SMA-parietal circuit is crucial for generating the sense of agency but it seems that the parietal cortex is the essential element. Patients with selective lesions to the SMA lose awareness of the “urge” to move . This is not surprising since the SMA is active

immediately prior to movement. Parietal areas are implicated in the sense of agency *per se* since they are essential for the monitoring process and are modulated by the SMA.

Blakemore (Blakemore et al., 2003, Blakemore and Sirigu, 2003, Blakemore et al., 2002) suggests that automatic motor control uses cerebellum-based prediction whereas controlled processes use a prediction system based in the parietal cortex. However rather than speaking of the prediction being located "in" a particular circuit it might be more helpful to think of the predictive systems as being distributed across the respective circuits precisely because any action involves or potentially involves different components of the system. When we perform an automatic action like walking down stairs the cerebellum modulates the parietal cortex. When we perform a controlled, visually-guided action, like walking through a minefield, the activity of the parietal cortex is modulated by the SMA under prefrontal direction as we slowly and carefully integrate visual reafference with proprioceptive predictions.

Lafarge and Franck summarise evidence from lesion and imaging studies of motor control and their relevance to the loss of sense of agency in schizophrenic patients. "A corollary of the motor signal involved in intended effort, in the SMA, could be sent to attenuate parietal activity due to proprioceptive feedback during active and so effortful movements. The balance between both phenomena might be abnormal in schizophrenia" (Lafarge and Franck, 2009 p284)

Awareness of Predictions in Schizophrenia

This account predicts that schizophrenic passivity experience should *not* be characterized by problems with automatic control, and passivity experience should be associated with unusual activation in parietal networks involved in controlled processing. Activation in inferior parietal areas which would otherwise be attenuated by properly functioning

prediction circuitry is not attenuated in schizophrenia (Spence, 2002, Danckert et al., 2004; Maruff et al., 2003)

In an important experiment Fourneret and Jeannerod had subjects trace a path from their body midline to a target directly in front of them. The subject's view of their moving hand was occluded until the final 30% of the movement. For the first 70% patients saw a computer-generated trace of the movement path. On some trials the experimenters introduced a deviation of 15% into the movement path so that if uncorrected the trace would veer off to the right. Subjects were able to compensate for the perturbation during the occluded section of the movement, with the result that when the hand came into view, the hand was to the left of the midline. So automatic control does not depend on explicit visual feedback but must use the automatic motor loop measuring body feedback against a prediction.

And in fact in Fourneret and Jeannerod's experiment schizophrenic's *automatic* performance was not different from that of neurotypical subjects (Jeannerod et al., 2003). They were able to correct for the perturbation in the absence of visual feedback. This concords with other experiments which test automatic control in schizophrenia. For example online schizophrenic motor control, like that of neurotypicals, obeys Fitts law, which says that duration of a movement is a function of its difficulty(i.e. there is a speed-accuracy trade off) Similarly schizophrenic patients show normal visuomotor adaptation to deviating prisms, another test of online visuomotor control (Danckert et al., 2004, Saoud et al., 2000, Decety and Lindgren, 1991).

And in fact versions of motor control tasks show that frequent visual feedback "reduced motor error and *differentially modulates the neural activation related to visuomotor processing in the cerebellum, parietal cortex and premotor cortex*" (Vaillancourt et al., 2003). They also suggest that this circuit dissociates from the circuitry involved in more automatic motor control. This finding is consistent with now numerous experimental and lesion studies suggesting that automatic motor control does not require the explicit representation of visual information.

Reviewing this literature Danckert et. al. express the consensus when they say "on-line monitoring and adjustment of action [automatic processing] is unaffected in patients with schizophrenia" (Danckert et al. 2004 p253).

So the experience of passivity is more likely to arise in deliberate, visually guided, control of action. Fourneret and Jeannerod's (2003) experiment described above provides a way to test this idea, because the last 30% of the movement is not occluded and when the hand appears it is fifteen degrees to the left of a straight line to the target. Neurotypical subjects explained this unexpected visual information by attributing the discrepancy to the computer indicating that they were able to become aware that they had intended a different movement than the one they actually made. Schizophrenics with positive symptoms did not, leading to the conclusion that "online control can coexist with a tendency to misattribute the source of error." p253.

A series of further experiments pursue the idea that passivity experience arises when predictive circuitry misfires during deliberate control of action. For example Mlakar et. al. asked schizophrenics to deliberately manipulate a joystick. They were able to do so correctly but reported vivid passivity experiences (Mlakar et al., 1994). This suggests that the automatic components of the action hierarchy were unaffected and that the sense of agency is generated when the action becomes the object of higher level processing. Spence et. al. pursued this paradigm in the course of a PET study with the same result (Spence et al., 1997b).

Daprati *et. al.* (Daprati et al., 1997)performed the following experiment. A Subject's hand and an Experimenter's hand, covered by identical gloves, were filmed by two different cameras. The subject performed actions (extending and retracting fingers etc) and viewed the image of the action performed on a television screen. From time to time, the image of the Experimenter's hand was substituted on the screen, sometimes making the identical movements, sometimes different ones to those made by the subject. The subject had to say whether the movements on the screen were those of his or another's

hand. The situation is the one described Pacherie, and probed by Fourneret and Jeannerod, in which close attention to visual information in the process of visually guided action should make the experience of authorship salient.

The two "easy" cases are firstly, those in which the image on the screen is that of the subject's hand, and secondly in which the image is of the experimenter's hand performing a different movement to that made by the subject. In the first case the image was congruent with the prediction of reafference and in the second case it was not. In both cases subjects correctly judged whether they were the author of the movement. However in the difficult case, in which the experimenter's hand made the same movement as that of the subject, subjects misattributed the movement to themselves in about 30% of a cases.

In the difficult cases there is no mismatch between the predicted movement and the visual information provided by another source. Hence the agent incorrectly attributes the action to herself. In this respect the case resembles the phantom limb case in which the patient received false visual information *confirming* a prediction.

Constrast this case with that of *discrepancy* between prediction and outcome. In such cases we become aware of trying to execute a movement that is different from the one we are actually executing because the predictive activity in the parietal cortex is not attenuated. This is why the "easy case" in the Dapriati experiment is one where the Experimenter makes a different movement from the subject.

One striking feature of Dapriati's experiment, and others with similar structures, is the fragility of the ability to determine, on the basis of visual feedback, whether an action is produced by oneself. In the difficult case, in which subjects saw the Experimenter's hand make the same movements as those they intended, 30% of the subjects misattributed the movement to themselves. This is consistent with the idea that visual information tends to take precedence if there is a conflict.

Daprati *et. al*. also tested two groups of schizophrenics, one experiencing delusions of influence, and the other not, on the Experimenter/ Subject Same/ Different task. The error rates on the difficult task (In which the experimenter's hand made the same movement as that of the subject) were 80% for the delusional and 50% for the non-delusional subjects. (Recall that the error rate for normal subjects on the difficult task was 30%.) Both groups had error rates of between 1% and 7% on the other "easy" tasks.

In another experiment subjects were asked to grasp a joystick and make simple movements, straight ahead to the right or the left (Jeannerod et al., 2003). Feedback was provided by a screen which displayed an image of a virtual hand. As in the Dapriati *et. al.* experiment the image on the screen was manipulated by the experimenters so that, in some cases, the movement of the virtual hand was identical to that made by the subject and in others different. The variation was introduced either by delaying the movement of the image by different intervals or by introducing a different angular bias. Normal subjects misattributed the movements to themselves when the delay was less than 150ms and the angular bias less than 15 degrees. At greater interval of delay, or if the angular bias of the image varied by more than 15 degrees, normal subjects correctly said that the movement was *not* the one which they had initiated.

This experiment tells us the about the threshold of visually represented temporal and spatial discrepancy to required to induce errors or attribution. At 300ms delay all 24 schizophrenic subjects responded at chance. However there was an interesting difference between influenced and non-influenced patients on the spatial test, in which subjects were confronted with angular deviations from their intended movement. In these cases the threshold (15degrees) at which non-influenced schizophrenic performance degraded was similar to that of normal subjects, whereas influenced patients continued to respond at chance until 30 degrees of variation. Thus Jeannerod *et. al.* concluded "only influenced patients were impaired in attributing movements with angular biases."(p.12)").

These cases substantiate the idea that schizophrenic passivity experience arises, as Pacherie suggests, when action is being controlled using explicit visual information. Passivity experience results from the failure to attenuate activity in the parietal cortex, especially inferior parietal cortex, essential to the integration of visual and bodily information *in controlled processing*.

This hypothesis is consistent with another finding. In normal subjects Fitts law applies to automatic and controlled movements *including imaginary movements*.(Danckert et al., 2002) When the subject imagines moving the duration of the process is a function of the difficulty. As we saw above imaginary movements generate the sense of agency since activation in parietal circuitry is artificially maintained by the absence of reafferent feedback. This is further evidence that the sense of agency is generated in the process of controlling action from above, since an imaginary movement is entirely under voluntary control.

However in schizophrenic subjects with positive symptoms imaginary movements do not obey Fitts law. This suggests that schizophrenics are not actually in control of the imaginary movement once it is "launched". For them the imaginary movement is experienced as passive because they do not produce the typical pattern of parietal activation which accompanies high level control (Danckert et al., 2002).

In a very elegant study Blakemore et. al. hypnotized subjects whose arms were attached to a pulley apparatus and gave them two instructions. In the first they were told to raise their arms and in the second that the pulley would raise their arms. The pulley did not actually exert any force. Highly hypnotizable subjects moved their arms in response to both instructions but in the second case they reported no feeling of agency, attributing the movement to the pulley. In effect hypnosis induced the passivity experience for controlled action characteristic of delusions of alien control. PET imaging showed

increased activation in the parietal cortex and cerebellum in cases when the subject attributed the source of the movement to the pulley (Blakemore et al., 2003).

Blakemore treats this as evidence that predictions available to consciousness are represented in the parietal cortex and automatic predictions unavailable to consciousness in the cerebellum. This is consistent with findings that patients with cerebellar but not parietal lesions have difficulty with automatic control of action but not with generating a sense of effort when they attend to those actions (Sirigu et al., 2003). Thus Blakemore summarises the consensus view in these cases that awareness of self-initiation of action depends on attenuation of activity in the right inferior parietal cortex by the SMA. "the prediction made by the parietal cortex is concerned more with *high level prediction such as strategic planning actions*" Furthermore she suggests "Perhaps the predictions made by the parietal cortex can be made available to consciousness" p243

In the normal case it seems that experience of agency is a subtle accompaniment to the visual control of action and recedes from attention if we attend exclusively to the visual stimulus. Perhaps schizophrenics pay too much attention to incoming visual information with the result that visual information becomes too salient at the expense of the sense of agency. This might arise because visual information is not properly gated, because frontal systems are deprived of resources or because of hyperactivity in the parietal regions Jeannerod et. al. (2003) found in a visual monitoring task that baseline levels of parietal activity were higher in schizophrenic patients than controls). Given that activity in parietal regions depends on modulatory activity in distributed circuits in which it participates there might be no single cause for hyperactivity (Danckert et al., 2004).

To summarise, schizophrenic subjects seem to have difficulty generating the sense of agency that normally arises when a person switches to controlled monitoring of action. Controlled monitoring normally involves inhibiting activation in parietal areas and schizophrenics have abnormal activation levels in these parietal networks. That higher levels of parietal activation are accompanied by loss of a sense of agency is most vividly demonstrated in Blakemore's hypnotic induction of passivity experience in neurotypical

subjects. That experiment and many others suggest that passivity experience can be produced from the top down, by goals or expectations explicitly represented by failure of attentional processes, from the bottom up by abnormalities intrinsic to parietal circuitry or distributed systems in which it participates such as the cerebellar-parietal or SMA-parietal circuits. The final cause of passivity experience however seems to be hyperactivation in the right inferior parietal cortex.

Passivity and Externality

If this is a plausible explanation of passivity experience in schizophrenia we still face the additional question: "why does the schizophrenic attribute authorship to someone else?" Why is passivity experience not simply reported as a felt loss of agency but *externalised* as the delusion of alien control?

Interestingly however there are some neuroscientists and philosophers who do not find this question relevant. For them the experience in question is of externality.

An eloquent statement of this view is given by Shaun Gallagher:

> the attribution of agency to another is not the result of a misinterpretation or cognitive accounting that would force the subject to infer that he is not the agent, someone else must be, or a supplemental account generated by introspection, the odd result of a productive narrative; rather *it is a genuine result of what is truly experienced* (Gallagher, 2004).

If this is the case, and we accept the account of the neural correlates of the experience then it must be the case that hyperactivation in the right parietal cortex signals, not just that an experience was unintended but that *someone else intended it*.

One attractive feature of this view is that it would explain the externality of schizophrenic experience without the need to invoke a further cognitive or metacognitive process. Non-attenuated parietal activation would automatically signal externality due to its role in a system that distinguishes between agents as causes of an action. As developed by Lafargue and Franck this idea suggests that different patterns of activation in the parietal cortex encode information about the author of the action.

> At the level of phenomenal consciousness such switches between attractor states could lead schizophrenic patients to experience at the same time a lack of control of some of their actions (including a lack of awareness of willed effort) with an intense feeling of agency for those actions thus creating conditions for the so called delusions of control. The hypothesised instability of the " who network" and its subjective counterpart in schizophrenia might result from abnormal SMA and parietal activity (Lafargue and Franck, 2009 p286).

This idea is nicely captured in Jeannerod's interpretation of the experiment that presented subjects with distorted visual feedback of their own movements (Jeannerod et al., 2003). They found that in neurotypical subjects parietal activity increased proportional to the degree of distortion when subjects saw a distorted image of their own movement. This is consistent with the idea that parietal activity is higher for unpredicted movements. In schizophrenic subjects not only was the baseline level of activity in the inferior parietal lobule higher than in neurotypicals but it was not modulated by the degree of distortion. As Jeannerod put it "the patients really had no cues (as inferred from the change in activity in the parietal lobe) about whether they saw their own movements *or those of an alien agent*" (Jeannerod 2006 p98).

On the face of it however the patients really had no cues about whether they saw the movement initiated by their motor cortex. Nothing we have discussed so far suggests that the parietal cortex attributes external agency to unpredicted movements. It merely generates the sense of agency for controlled processing, a sense which can go missing in schizophrenia.

Blakemore gives a similar interpretation of induced passivity experience in her hypnosis experiment. When they experienced passivity as a result of hypnosis subjects automatically attributed authorship to the pulley. Blakemore treats this as an example of the *experience of externality* not the experience of passivity subsequently attributed to an external author. If this is correct a high level belief or goal can change the structure of experience itself by introducing representation of external control into the system that produces experience.

However this looks like a case in which the experimental contexts establish the default assumption that "if I did not initiate this movement someone/something else must have". However from what we have seen so far the modulation of parietal activity only tells the subject whether or not she initiated an action, not whether someone else initiated it

In thinking about this issue the disanalogies as well as the analogies between the way predictive coding works in perception and action are important. In the case of perception detection of discrepancy between predicted and actual state *automatically signals externality*. If I move my eyeball five degrees but retinal displacement is ten degrees it is because the object has moved five degrees. Perceptual systems evolved to compute this automatically. The only cases in which this is not the case are extremely artificial ones such as manually moving the eyeball. Not surprisingly these cases are experienced as movement in the external world, but this is an illusion generated by the non standard origin of the retinal displacement which produces a representation of externality when not predicted.

If I try to move my arm ten degrees but it only moves five degrees the discrepancy does not automatically signal that *someone else* has moved my arm five degrees. It signals error due to noise in the motor control systems or unexpected interference. (Perhaps muscles in my arm are fatigued from participating in motor control experiments, leading

my SMA to tell my parietal cortex to expect more movement than my muscles can generate).

Evolution has not posed us with the problem of determining which movement are ours *rather than someone else's*. It has posed us with the problem of determining which of our movements are consequences of motor commands in order to compute and resolve error. Thus there seems no reason to think that the motor system would need to use predictive coding to disambiguate the agent of an action rather than to simply control our own action. Thus in the case of action forward models need to distinguish afference from reafference to *control movement*, not to attribute authorship of movements.

Mirror neurons and other bodies

Or so it seems. Jeannerod relies on recent findings to show that activation in parietal circuitry *can* be produced by the observation of other's actions. Experiments with macaques provided the first evidence. Single cell recordings from the premotor cortex were obtained both while they performed actions, such as hand and mouth grasping and observed those same actions being performed. Experimenters found that cells in the ventral premotor cortex (F5), baptised "mirror neurons", fired in both conditions. These cells respond to both motor and visual representations of actions. These cells also fire when monkeys see a transitive action whose target is occluded which suggests that the representation of movement as intentional is encoded "upstream" to the mirror system. The mirror neuron system is not part of an action categorisation system but responds to movements *already* categorised as intentional, as one would expect of predictive circuitry (Gallese et al., 1996, Gallese and Goldman, 1998, Gallese et al., 2004, Iacoboni et al., 2005, Oztop et al., 2006). The functional significance of mirror neurons for macaques is not yet clear. They have been interpreted as playing roles in imitative learning and communication but there are difficulties with these interpretations (Gerrans, 2010b, Gerrans, 2010a)

Like macaques humans appear to have mirror systems which are activated by action, observation *and imagination*. These systems include the parietal predictive circuitry. Interestingly in humans the range of actions which elicit mirror activation is wider than macaques. Intransitive actions (not directed towards a manipulable target) and pantomime actions as well as transitive actions produce mirror activation. The controversies about functional interpretation increase in the human case with roles for mirror neurons being proposed in social cognition and communication, language and motor learning (Gallese and Goldman, 1998, Goldman, 2006, Oztop et al., 2006, Gerrans, 2010b).

Leaving these controversies aside, the presence of mirror neurons suggests that parietal activation, which in the forward model signals that an action is not self-initiated, can be produced by the observation of actions performed by others. In a way however this is not surprising. Observation of *any* action will produce parietal activity.

The best explanation of this is that in order to monitor action we need to represent it both egocentrically and allocentrically. That is to say from the neutral third person perspective as well as in relation to ourself. The fact for example that I missed my grasping target by 2cm can be represented in two ways. Egocentrically, in virtue of the fact that *my* arm extended 2cm further away from *my body* than *I intended*. Or by representing the distance between arm and target: 2cm. The latter representation is allocentric.

Perceptuo-motor control requires the integration of both types of information. In terms of the forward model for example vision tells us that the arm is 2cm from the target, which together with the goal (reach to the target) creates the inverse model. This must be transformed into a motor command: move 2cm. That motor command creates a prediction sent to the parietal cortex which modulates its response to incoming visual information. That incoming visual information *is allocentric:* in this case 2cm movement of the arm. When the arm moves 2cm the parietal cortex does not respond as strongly as if the movement was 4cm because response to that particular movement has been anticipated. This is simply to repeat that prediction is the attenuation of parietal responses.

Of course at the level of explicit visual experience the information is egocentric in the sense that it is experienced from the subject's perspective. However the SMA-parietal-frontal circuit does not care " whose" movement matches or mismatches a motor intention.

An interesting question which is relevant here is how we distinguish intended from non-intended movement in other people. The parietal cortex does not seem relevant to this task. People with parietal lesions have problems with action control but not with determining whether other people are walking or stumbling, or whether a defender's hand-ball is intentional. The reason is that we use specialised action detection systems for these tasks. In particular regions of the superior temporal sulcus respond preferentially to goal directed movement (in both humans and macaques). The STS does not encode motor information but information that movement is goal directed.

Thus when we observe another person's action we already have the information that what we are observing is an action. Also the very fact that we are observing another person tells us that the action is theirs. It seems then that we would not need to recruit the parietal cortex to tell us that an action is being performed by someone else *in virtue of the fact that we did not initiate the observed movement*.

A recent imaging study parsed components of the distributed circuitry required in action control (Lestou et al., 2008) The experimenters summarised their findings this way:

parietal regions and the superior temporal sulcus process the perceptual similarity between movements and may support the perception and imitation of *abstract action goals* and movement styles. Thus, our findings propose that fronto-parietal and visual areas involved in action understanding mediate a cascade of visual-motor processes at different levels of action description from *exact movement copies to abstract action goals* achieved with different movement styles.

I take this to mean that action can be understood both in terms of ends (superior temporal sulcus) and means (parietal cortex). We understand an observed movement as an intended movement mainly because of the superior temporal sulcus. As Gergeley Csibra puts it "STS neurons seem to 'understand' actions quite well, and it is plausible to assume that they send (via PF) pre-processed signals about actions to the premotor areas that include information about the goal or the meaning of the observed action" (Csibra, 2005). The parietal cortex responds to the movement by detecting whether or not it was predicted (*exact movement copies*) not the goal or the agent.

To say that the parietal areas represent movements is to say that their activity is modulated by the perception of movements. Non-attenuated parietal activity *signals that the movement was not intended*. It only follows from the fact that a movement was not intended *that it needs to be corrected*, not that it was intended by someone else. The difference between attenuation and non attenuation does not represent different agents of an action: it merely represents whether or not the movement is the outcome of the observer's motor instruction (Oztop et al., 2006).

Of course there might be cases when we observe a movement and there are no other cues about the agent: but these tend to be artificial, not to say experimental (the equivalent of manually moving an eyeball or paralysing it with curare), rather than recurring adaptive problems for which we would have evolved a specialised solution. In general then I conclude that parietal activation is not specialised for determining *who* intended the action. Rather it determines for any movement whether it is a consequence of a motor instruction. It evolved to control movement not identify the agent.

Another advantage of this interpretation is that it can economically explain the opposite cases which also sometimes occur in schizophrenia. For example one patient during the second world war used to watch planes flying overhead and feel that she controlled their trajectories. The fact that these feelings can arise for movements of inanimate objects suggests that what is happening is just the attachment of a sense of agency to a perceived movement.

Passivity of Thought.

So far we have concentrated on the passive experience of action. The forward model also provides an attractive but more complicated way to explain delusions of thought insertion. If delusions of thought insertion involve the passive experience of thoughts attributed to an external agent then the structure of processing involved may be similar. This would require us to reconceptualize thinking "as a motor process" in John Campbell's phrase (Campbell, 1999).

Extending the predictive coding approach to the case of thought faces problems. The first is conceptual: actions realise a goal so the idea of predicted action makes sense. But we do not typically intend to reach a specific goal when we think. Consequently we do need to intend or predict thoughts. If the thought occurs to me to go to Japan next summer I do not suddenly become aware that that is not the thought I intended to have. Whereas Dick Cheney *did* experience a mismatch between intention and outcome when shot his hunting companion instead of a duck. At face value the process of thinking is quite dissimilar to the process of acting.

The second set of problems arises from the fact is that there is no clear consensus on the phenomenology of thought insertion, the cognitive processes involved or the way to interpret the sparse neuropsychological data. For example clinicians recognise a distinction between auditory verbal hallucinations (hearing voices) and inserted thoughts. Clearly in both cases the experience is self-initiated but there is as yet no principled basis for distinguishing them (David, 2004).

One possibility is that fMRI or another imaging methodology might help distinguish cases such as this For example it might be the case that auditory hallucinations exclusively involve endogenous activity in the auditory system. However the fact that the hallucinations are of voices suggests that things are not as simple as endogenous auditory

activity. Why not hallucinations of innocuous noises? Furthermore the voices in auditory hallucination are not always sensed as originating from beyond the mind of the agent.

It is for this reason that some theorists have proposed that we treat both auditory verbal hallucinations *and* inserted thoughts as passively-experienced episodes of inner speech (Jeannerod et al., 2003, Langland-Hassan, 2008, Shergill et al., 2003). On this view the subject generates an episode of inner speech but is not aware of doing so *and* attributes it to another agent. Note that this form of explanation requires at least two steps. The first is to explain how it is that a patient might experience a self generated subvocalisation as not self-initiated, and the second is the attribution of that experience to an external author.

This approach solves the conceptual problem at the price of narrowing the target phenomenon to the case of inner speech. The problem here is that is makes delusions of thought insertion *not actually cases of thought*. The process of thinking is the process of moving from one thought to another: association, deduction, induction, analogical reasoning, imagery, pattern completion and so on. Speaking, while it might express the results of these processes, is a different cognitive process.

However this may not be a decisive objection to treating delusions of thought insertion as inner speech. There is an intimate connection between thought and language. On some views natural language *is* a medium of thought (Carruthers, 2002), on other views language plays a coordinating and indexing role, keeping track of thoughts while they are manipulated. Inner speech might be an important thinking *tool* (Vygotsky, 1986, Vygotsky, 1978, Jones and Fernyhough, 2007a, Jones and Fernyhough, 2007b). The view that thought itself is non linguistic is consistent with the view that we often experience the *outcome* of thought processes as episodes of inner speech. Not only that but we often initiate thought via inner speech. Perhaps inputs and outputs to thinking are encoded linguistically even if thinking is not.

Thus thoughts *are* intimately connected with inner speech and many of the outputs of thought processes are experienced as inner speech. It is this class (and I suspect only this class) of mental states that are candidates for explanation using the forward model.

If we treat speech as a form of action and the thoughts of thought insertion as episodes of inner speech then the same structure of explanation may apply to passively experienced actions and passively experienced inner speech. One important difference is that actions are overt, that is actually produced and observed. Inner speech remains inner. However Marc Jeannerod's discussion of action and imaginary action suggests a way to approach this problem. On his view inner speech is a form of *covert* action. It stands to speech as imagining an action stands to action.

As with normal actions covert or imaginary speech should exploit the same control system as normal or overt speech. In both cases the overt component is inhibited with the consequence that the relevant predictions are not cancelled. Thus in the normal case inner speech, like imaginary action, should produce a sense of agency as activity in predictive circuitry is not attenuated. The attempt to make a phoneme should produce a corollary discharge that attenuates activity areas which process feedback consequent on speech production (Ford and Mathalon, 2004, Ford and Mathalon, 2005, Ford et al., 2007). (Allen et al., 2007, Jones and Fernyhough, 2007a).

Just as schizophrenics are poor at monitoring imaginary actions perhaps they are poor monitors of inner speech *precisely because inner speech is a form of imaginary action*. If this is the correct explanation of delusions of thought insertion then we would expect to see abnormal activity in brain areas which monitor covert speech. And here are some imaging studies which suggest that production of speech does attenuate activity in speech processing circuits in both *talking and thinking*. These studies also found that this attenuation is absent in schizophrenics hearing voices (Blakemore et al., 2000, Ford and Mathalon, 2005, Ford et al., 2007, Heinks-Maldonado et al., 2007, Allen et al., 2007, Jones and Fernyhough, 2007a).

A complicating factor is that a lot of this work does not discriminate between auditory verbal hallucination (AVH) or "hearing voices" and "inserted thoughts". In equating AVH and the representations that trigger delusions of thought insertion with inner speech I am taking a position in a debate which is not yet conclusively resolved (Jones, 2010, Jones and Fernyhough, 2007a, Jones and Fernyhough, 2007b). However the evidence is tending in the direction that AVH, like inner speech, involves circuits which monitor speech rather than simply responding to auditory signals.

Even on the assumption that inserted thoughts are episodes of inner speech, unaccompanied by a sense of agency, the same problem arises as in the case of actions. The forward model hypothesis only explains passivity not externality.

However the default attribution of externality might be easier to explain for speech than passively experienced bodily movement. Evolution has not posed us the problem of determining which bodily movements are ours. However the problem of disambiguating speech signals, firstly from the noisy background and secondly from the speech of other conspecifics is an adaptive problem for a social species like ours. When acquiring language we might need to be particularly attuned to subtle differences between our actual and intended vocalisations. Consequently if we lose the ability to monitor our vocalisations the default attribution of externality might spring to mind automatically.

External Attribution of Thoughts

There are other accounts of external attribution of thoughts which address this problem but ultimately they do not seem able to explain the passivity of *experience*. For example, when John Campbell (2001) proposed that we think of thinking "as a motor process" he developed an account of the externality of inserted thoughts. Campbell distinguishes thoughts from speech and, unlike the account developed here, focuses on thoughts.

He thus directly faces the problem, avoided by focusing on speech, that we do not intend to think thoughts, predict which thoughts we will think as a consequence and match actual to predicted thoughts. The problem is not just that we are not aware of these computations (we are not aware of these computations for automatic actions but they do occur) but that there seems to be no principled reason to postulate them. In the case of actions, including speech, some monitoring and control mechanism is required and the forward model is a plausible candidate. In the case of thoughts there is no reason to think that such monitoring is required. There is no sense in which we need to evaluate daydreams for successfully meeting our goals for example.

Nonetheless as the idea of a "train of thought" reminds us thinking is not random, it has a theme or a goal around which thoughts are integrated, however loosely. Even daydreams have topics that focus the associative process. John Campbell expresses this as the ability to keep thoughts "on track" and suggested that it is in this respect that thinking is monitored.

The difficulty as we saw above is how to make idea more precise without postulating intentions to think about certain topics as keeping the stream of thoughts on track. Campbell realises that there is a problem here and suggests that while the stream of consciousness is not under our volitional control in the same way as our actions, there is an important, analogous, sense in which thoughts are our own. That is the sense in which what we think is consistent with our background psychology. Background psychology can be thought of the total of dispositional psychological facts (beliefs, desires emotions and psychological traits) that form a more or less integrated unity and allow us to think and behave consistently over time. As he says "what makes my occurrent thoughts mine is *not* just that they show up in my stream of consciousness. What makes them mine is, in addition the fact that they are the products of *my* long standing beliefs and desires and that the occurrent thinking can affect the underlying states." (my italics)

The question raised is how we detect this "mineness". One obvious way is to check for consistency with other states. Is an occurrent episode consistent with other emotions, memories beliefs and desires? The notion of consistency involved need not be full-blooded rational consistency but the looser notion of autobiographical coherence. The self has a psychological history which it is constantly updating revising and reinforcing. Some episodes of thought might arise in a mind and very hard to accommodate. A sudden burst of psychotic anger in the mind of a person whose habitual response to the world is tranquil reflection is hard to incorporate.

This type of autobiographical coherence is not the type of consistency Campbell has in mind. Campbell is trying to explain how a thought might be *experienced* as external not subsequently attributed to an external author on the basis of incoherence with autobiographical precedent. His idea is that when a thought is produced something like an efferent copy is produced by the neural mechanisms involved. That copy remains submerged in the subpersonal machinery but is matched to the consciously experienced thought. In effect this makes a consciously experienced thought a reafference of an unconscious process. When the efferent copy goes missing or the matching fails we experience a thought as not self-initiated.

However even if there were such a process it seems it would not play the role of making a thought coherent with background psychology. Tracking the fact that *we* initiated the thought won't stop the problem of irrelevance or incoherence. We can, and do, initiate incoherent thoughts.

Secondly Campbell's proposal in fact would only solve the problem of passivity, not external attribution, just as loss of access to predictions in the action case does not produce externality. And it seems that we can far more economically explain the sense of passivity for inner speech. The price is of course that speech is not (always) thought.

External attribution and psychological coherence

Why then are passively-experienced episodes of inner speech attributed externally? There may be a clue in Campbell's argument that thoughts are tracked for psychological coherence. Perhaps some passively experienced episodes of inner speech express thoughts that do not cohere with background psychology and are thus subsequently attributed to another person. This idea is also shared by motivational accounts of delusion which argue that the reduction of distress is a central contributing factor in some entrenched delusions. Rather than acknowledge a distressing thought or emotion the subject attributes it to someone else.

The problem of psychological coherence is a genuine problem and there is good evidence that people do strive to produce a coherent autobiographical model. We noted earlier the role of the ventromedial prefrontal cortex in making thoughts essentially indexical by making their automatic emotional and motivational associations salient. Not only that but there are numerous cases where people incorporate false beliefs and memories into their autobiographical narratives on the grounds that they cohere emotionally and contextually with the preexisting autobiography (Mitchell and Johnson, 2000). Together these facts suggest that we actively try and reduce psychological friction or ego dystonia.

As we saw in the discussion of narrative psychological coherence involves temporal and emotional coherence as well as dispositional traits to think in distinctive patterns. Thoughts do not arise in consciousness out of a vacuum but as a product of the interaction of occurrent experience with long term memories organized as an autobiography. Thus a good way to put Campbell's point is to say that our thoughts, insofar as they are distinctively personal, cohere with an autobiographical narrative.

However establishing narrative coherence is not the same problem as the problem of establishing for any thought that we rather than someone else produced it. The processes

of narrative construction or reconstruction are subsequent to the thought. Thus any attribution or misattribution which arises as a result of that process is a metacognitive response to experience.

Stephens and Graham address this problem in their account of inserted thoughts. They say that some thoughts, though inconsistent with background psychology, may nonetheless "seem to betray an agency, and intelligence, that accounts for their coherence, salience and directedness. Thus she may have the strong *impression* that someone is thinking in, or speaking to her." (Stephens and Graham, 2000). The idea seems to be plausible at the phenomenological level. To take an example of Stephens and Graham, a distraught and sleepless parent may experience the impulse to hit a chronically crying child. Such a thought is entirely inconsistent with the overall pattern of her attitudes. Thus the thought is external to her background psychology in the sense described by Campbell.

Stephens and Graham follow a suggestion of Hoffman's (1986) that experiences which are misattributed are *ego dystonic*. That is inconsistent with the subjects conception of herself. They note that the "voices" misattributed by delusional subjects are often suggesting actions with which a subject might not identify. Suicidal or violent impulses are examples. The attribution of these thoughts to another source might remove distress caused by acknowledging a desire which runs counter to the narrative a subject has constructed to integrate her experiences.

For Stephens and Graham the externality of a thought derives from its being appraised as inconsistent with an autobiographical narrative: on the face of it a higher order and, possibly, dissociative process. It is not clear how *appraising* a thought as alien would generate the phenomenology of passivity or externality. Stephen's and Graham's use of the term *impression* suggests a distinctive phenomenology following from an appraisal of a thought as external, but it does not follow from the fact that a thought is appraised as

bizarre, unanticipated or out of character and that it *feels as if someone else is thinking it via your mind*.

There is a genuine dilemma here. Motor control theories can explain the experience of passivity, if not externality, but they do not apply to thoughts. Hoffman and Stephens and Graham can explain the external attribution in terms of motivated appraisals of ego dystonic episodes of thought but they cannot explain the experience of passivity.

Passivity of experience, externality of thought.

As we noted earlier a solution to this dilemma is to reconsider the relationship between inner speech, understood as a covert action, and thought. While language is not necessarily the medium of thought it does play a crucial role in tracking, indexing, and manipulating thoughts. This is why so much of our inner speech is planning, commentary and review, seamlessly integrated with other forms of thought such as offline rehearsal of action. The folk psychological equation of language with thought only creates difficulties if treated as the essence of thought rather than something intimately connected with it in our experience.

If inner speech is a form of covert action intimately connected with thought the passive experience of thought insertion can be explained in the same way as passivity of action in delusions of control. i.e. the sense of authorship is lost as a result of faulty activation in predictive circuitry. The result is inner speech unaccompanied by a sense of authorship and, since inner speech and thought are so closely tied together, the experience is indistinguishable in introspection from passive thought.

Of course this does not explain the externality of the delusion: the idea that another agent is inserting thoughts or using the subject's mind as a medium for thought. The experience of thoughts without thinkers, like actions without authors, triggers a default thought in

which experience of passivity is attributed to an external author. Once that attribution is made it functions as an expectation against which future experiences are appraised for autobiographical coherence and becomes more firmly installed.

Why does the default thought attribute an agent? Richard Bentall while not directly addressing this question suggests that the experiences in question are personally significant (Bentall, 2010). The voices may be accusing, threatening, highly distressing or sometimes encouraging but they are very rarely neutral commentary. This fact about them together with the experience of passivity may be sufficient to produce the attribution of agency. Bentall gives the example of a patient who constantly spoke to herself, recalling the encouraging words of her psychologist as she reacclimatized to life beyond the ward. The patient was behaving like a musician who repeats to herself the teacher's instructions while playing a difficult piece. As a result Mr Pinkerton's consoling words provided the soundtrack to her life. When her passivity experience recurred and she lost the sense that it was she herself who was producing the soundtrack the thought that Mr Pinkerton was still encouraging and guiding her seemed inevitable.

LOUIS SASS AND THE SCHIZOPHRENIC LIFEWORLD

Phenomenological theories of delusion aim to make the lifeworld of the delusional patient intelligible, to cross that interpretative border described by Jaspers. Some phenomenologists have argues that one reason delusions seem so mysterious is the doxastic framework itself. Understood as empirical beliefs many delusions are inscrutable. Thus part of the phenomenological project is the development of an alternative conceptual framework for understanding delusion.

In this chapter I discuss one of the most interesting and subtle phenomenological accounts, developed by Louis Sass. Sass concentrates on a subset of schizophrenic delusions which are at the bizarre end of the spectrum compared to those we have discussed previously. Nonetheless his account of the delusional lifeworld in these cases is relevant to any attempt to develop a non-doxastic account. In fact I shall argue that the features of delusion that figure in his account as problematic for doxastic theories can be accounted for in terms of cognitive properties of the default network. If this is right it reinforces the more general point with which we started. If we describe the cognitive properties of the mechanisms which generate delusion in sufficient detail then the personal level aspects of the phenomenon become explicable. The focus of explanation shifts from questions inspired by the doxastic framework (are delusional subjects rational? Are delusions beliefs?) to the explanation of the processes which generate

delusion. Although, historically, phenomenologists have been sceptical of cognitive neuroscience a rapprochement is clearly required if cognitive neuroscience is to explain experience.

Schreber's lifeworld

Sass's approach to schizophrenia is organised around an extensive case study of the *Memoirs of my Mental Illness* by Paul Schreber (Sass, 1994). At various times Schreber, a German judge, exhibited the whole range of positive and negative symptoms of schizophrenia. He had bizarre delusions of "nerves" and "rays" taking over his soul and controlling his actions, committing "soul murder" in the process. He witnessed and performed "miracles", especially being changed, or turning into himself into, a woman. He had delusions of omnipotence (influencing events in distant corners of the universe) omniscience (a deeper comprehension of the nature of existence than anyone has ever had before) and even omnipresence, at times describing himself as coextensive with the universe.

As well as these symptoms of psychotic experience, Schreber also showed behavioural signs of psychosis, as when he would fall into immobility for hours, stare into the sun or at a particular object, or mutter incomprehensibly. In his *Memoirs* Schreber documents these and other episodes in a fanatically and fantastically detailed description of schizophrenic experience. On encountering passages like the following, Jaspers view (partly based on the *Memoirs*), is difficult to resist.

> souls in nerve contact with me talked of a plurality of heads (that is several individuals in one and the same skull) which they encountered in me and from which they shrank in alarm crying 'for heaven's sake-that is a human being with several heads' (Sass 1992 p243)

> I can put this point briefly: everything that happens is in reference to me… Since God entered into nerve contact with me exclusively, I became in a way for God the only human being around whom everything turns, to whom everything that happens must be related and who therefore, from his own point of view, must also relate all things to himself

> This completely absurd conception, which was at first naturally incomprehensible to me but which I was forced to acknowledge as a fact through years of experience, becomes apparent at every opportunity and occasion. For instance when I read a book or newspaper one thinks that the ideas are my own; when I play a song or opera arrangement for the piano, one thinks that the text of the song or opera expresses my own feelings (Sass 1994 p61)

> I have to add that the female characteristic which are developing on my body show a certain periodicity at increasingly shorter interval. The reason is that everything feminine attracts God's nerves. Hence, as often as one attempts to make the female characteristics which are evident on my body recede by miracle; the effect is that the structures which I call "nerves of voluptuousness" are pushed a little under the surface, that is to say are not distinctly palpable on the skin, my bosom becomes a little flatter, etc. But after a short time the rays have to approach again, the "nerves of voluptuousness"(to retain this term) become more marked, my bosom bulges again, etc. Such changes occur at present in as short a period as a few minutes. (Sass 1994 p123)

> I mainly sat *motionless* the whole day on a chair at my table, did not even move toward the window where by the way nothing was to be seen except green trees.

(Schreber's physician, commenting on his poverty said:

> This physically strong man in whom frequent jerkings of the face musculature and marked tremor of the hands were noticeable was at first completely inaccessible

> and shut himself off, lay or stood immobile and stared with frightened eyes straight ahead of himself into space. (Sass 1994 p37)

Sass offers an original account of the schizophrenic lifeworld to account for passages such as this. On his view schizophrenia is frequently characterised by intense absorption with sensory and affective qualities of experience in which the delusional subject takes experience at its face value without thereby *mistaking* it for a representation of external reality. The attitude to experience is not that of someone trying and failing to establish whether it is veridical. Nor is it the attitude of someone simply assuming the veridicality of experience. For Schreber, as for many with schizophrenia, experience had a fascinating and magical quality that compelled his absorbed attention. We could say that experience, which is normally transparent and direct, took on a different aspect for him.

In fact, as Sass, and many others, have pointed out, delusional subjects often engage in a kind of double entry book-keeping in which many or most experiences are treated as subject to normal evidential standards (which can in some cases be imported into delusional thought processes) but the delusional experience is exempted. If this is correct, "reality-testing" or source-monitoring accounts, which assimilate delusions to empirical beliefs based on hallucinations, misrememberings or misperceptions are mistaken, because the delusional subject has not in fact lost the ability to determine whether or not experience is veridical (Mitchell and Johnson, 2000). The delusional subject is often well aware that others, and she herself when taking an epistemic attitude to delusional experience, would judge a proposition which captured its content to be false. Sass quotes one of his patient informants:

> I've never really held my beliefs about Pepperidge farms and microwaves but they've always had a strong feeling of fear and aversion related to my belief that nothing exists- however I have acted for a long time as if these beliefs were unquestionably true… but I've always had a dimension of doubt about these beliefs and of course I realise how profoundly irrational they sound to other

> people… I would much prefer to believe that I am delusional rather than that all these magical events and processes are real ((Parnas and Sass, 2001 p79).

This type of insight is reminiscent of that of Alexander's Capgras patient, with whom we staretd, who noted that other people would find the idea of a replacement family, cloned overnight, "unbelievable".

Sass's aim is to elucidate this ambivalent attitude. He does so by considering the phenomenological predicament of someone who is utterly absorbed with the felt quality of experience rather than the external world. Such a person would *appear* solipsistic in the everyday sense: their world would be constituted their own experience.

According to Sass, Schreber lived on the borderline of solipsism, inclined to regard all events as internal to his mind. He pursues this idea via a consideration of Wittgenstein's discussions of solipsism which was largely aimed at we might call Cartesian theorists of experience. For Cartesians we do not perceive the world but intermediate entities, experiences, from which we infer the nature (if any) of the world perceived. There is no way to establish whether our experiences are veridical other than via their coherence with ideas gained as a result of other experiences, each of which is similarly dubitable. Sceptics draw the conclusion that there is a veil of ideas between the subject and the external world. Consequently we can never have knowledge of the external world on the basis of experience. All we experience is the stream of our own experiences. An idealist solves this problem by saying that the veil of ideas *is* the world.

Wittgenstein's objection to this line of thought is that to entertain the sceptical attitude that generates it, we have to adopt a very unrealistic attitude to experience. For a sceptic experience, normally taken to be of the world, is now taken to be of intermediate entities which stand between the subject and the world. Experience, rather than its object, becomes the target of attention; and the mind, rather than being focused on the world, is now focused on experience itself. To generate this sceptical attitude we have to redescribe experience, via the indirect theory of perception and the model of the self as a

detached scanner of experiences, in a way that is not consistent with the way it feels to live in the world. Rather than giving us insight into the nature of experience the sceptical argument estranges us from our everyday experience and from the world. This is why scepticism can only be entertained as an intellectual exercise not actually lived.

Sass and Parnas suggest that this kind of lived estrangement is the phenomenological essence of (Schreber's) schizophrenia and offers important clues to the lifeworld of delusional patients, which, if interpreted correctly, allow us to understand their perspective on experience (Sass, 1994, Parnas and Sass, 2002, Parnas et al., 2008, Sass and Parnas, 2003, Parnas and Sass, 2001).

Thankfully, for most of us, most of the time, our engagement with the world is straightforwardly mediated by our mental life. We hear, see, feel and think about, objects, including our own bodies. However whereas normal people can attend, via their experience, to the world or their bodies understood as objects in the world, the schizophrenic attends, with special absorption, to experience itself. For the schizophrenic experience becomes detached from its normal role in mediating encounters with the world and becomes, so to speak, the world. Not in an ontological but a practical sense. Subjective experience prescribes the limits of action and thought.

The self of solipsism is a curious thing. Given that there is nothing in the solipsist's world but her own thoughts, agency is reduced to nothing other than the experience of the flow of mental events. In one respect she is an omnipotent God, since every event is a product of her mind, in another she is hardly an agent at all, just a locus of experience. Everything that happens is a product of the solipsist's mind, but awareness of these events is simply a matter of instrospection, not perception. She is all-powerful, but in a virtual world—which is to say, in another sense, in no world at all. All the solipsist does is watch herself thinking. The result is a kind of omnipotent passivity.

Schreber's *Memoirs* express the same kind of passive but intense attention to experience itself and a reconception of the external world, including his own body as dematerialised.

Sass makes another point about solipsism relevant to Schreber and schizophrenic delusion in general. Everything a solipsist says must be true. After all she is only reporting what is going on in her mind, to which she has introspective access. Since, however, the notion of truth itself implies intersubjective verification, to claim an objective reality for solipsistic beliefs is an empty speech act. As Wittgenstein points out, the solipsist is in danger of realizing that his grand pronouncements of cosmic centrality amount in fact to little more than the tautology that "my experience is my experience".

If Schreber's condition is such as to induce a vertiginous solipsism then it would explain his sense of omniscience, "subjective certainty" (his own words) omnipotence, passivity, depersonalisation, of a world newly created by him ("miracled up"). Sass's interpretation of Schreber's consistent references to "nerves" and "rays" ("God's way of seeing") and of his sexual transformations are also fitted into this interpretative framework. Rays are mechanisms by which events are scrutinised by God, or by Schreber himself: they are manifestations of a kind of meta-awareness or hyperreflexive consciousness. Since his own mind is, effectively, the world, he can produce miracles just by noticing or attending to things. Nerves are the mechanisms of a more immediate form of experiential awareness. Schreber associates the passive aspect of his condition with femininity and the delusions of omnipotence with masculinity, hence his transformation from one to the other in the act of attending to his bodily experience.

On Sass's account, Schreber's schizophrenia produced a distinctive attitude to experience, which differs from empirical belief. In fact, insofar as Schreber is a solipsist it would make no sense for him to have beliefs, in the normal sense in which a belief is accountable to both external reality and intersubjective standards of justification (Bortolotti, 2005). Kant in fact was one of the first to make the point that even an

internally coherent world constituted by subjective experience would not generate beliefs as we normally understand the term. Hence on his view delusions represent “a play of thought in which he proceeds and judges in a world not shared by other people, but rather (as in a dream) *he sees himself in his own world*” .

It is this lack of commitment to public standards of justification which leads to adoxastic accounts of delusion. Sass emphasises that Schreber did not *believe* his delusions. He quotes Schreber himself explaining a delusion that his psychiatrist’s soul had lodged in his (Schreber's) body in the form of a cobweb "an actual or subjectively certain event... whether or not other people can believe me." (Sass p42). At another place he says "all this, naturally, only in my own imagination" (p27)

Cognitive Phenomenology

Sass’s phenomenological account of Schreber’s delusional state may not be a template for the understanding of every delusion. In particular, in monothematic delusion introduced by focal lesion, the lifeworld of the subject does not seem to be totally reconstructed in the same way. However, Sass’s account is very important since it is one of the only completely developed accounts of the delusional mind, which treats delusion as a *sui generis* psychological state not to be understood within the philosophical or psychological framework of empirical belief fixation.

In that respect it is very congenial to the general picture of delusion developed in earlier chapters which treats delusion as an attitude not best characterised in epistemic terms. We explained that attitude in terms of a state of the control hierarchy in which default systems, which create an autobiographical context for experience monopolise the mental economy. The mechanisms that create that monopoly may differ from delusion to delusion (lesion, dopamine dysregulation, abnormalities of connectivity or conductivity between of the default system and other components of the hierarchy) but in each case

the result is to focus metacognitive resources on subjective experiences and thoughts. Not only that but often the effect of dopaminergic enhancement of these experiences and default responses is to make then especially salient, reinforcing the halo of significance which seems to attach to delusional experience.

It is one attractive feature of this account that the delusion can coexist in a psychology otherwise capable of rational belief fixation. However that coexistence depends on the fact that those intact rational capacities are not directed at confirming or disconfirming the delusion. Rather, they are employed to construct the background or stage setting for delusion. Delusions stand to this background in the same way stories which involve magical or supernatural events take place in a background world whose mundane empirical and logical laws remain consistent. The psychological understanding of such stories depends on the interaction between imagination and belief required for the story and established by the context. Similarly default thinking governed by subjective patterns of association coexists in a normal psychology with more decontextualised forms of thought. Even the chap who thought he was inhabited by a lizard shared his beliefs about reptiles, scales, claws and cold-bloodedness with the rest of us. The madness of the delusion consists precisely in the effortless combination of fantastic and mundane.

Delusion represents an extreme case in which the default circuits which produce the subjective imaginative associations are hyperactive and unable to be supervised by decontextualised processes. This lack of supervision may be intermittent, due to neurotransmitter fluctuation, or permanent due to lesion to the necessary circuitry. Or it may be a result of entrenched patterns of referral and supervision produced over a lifetime. When we retrace the mechanisms of referral and supervision in each case of delusion we will discover that default circuitry activated by the referral of salient information is (hyper) active. At the same time circuitry required for decontextualised processing is typically underactive, compromising the ability to refer default thinking for supervision.

If something like this account is correct then Sass's phenomenological account can be underpinned by a cognitive theory which links the phenomenology to underlying neurobiology. Schreber, like other delusional subjects was a hostage to experience because he could not escape from default mode.

The aim of this book was not to vindicate Sass's account at the expense of other neurobiological or epistemic accounts but to preserve the phenomenology while showing how it might arise as a result of the influence of biology on cognition at all levels of the control hierarchy.

This bottom up approach has produced the same conclusion as Sass's phenomenological investigation. Namely, approaches which conceive of delusions as empirical beliefs can be misleading. The reason is that they direct attention to mechanisms of belief fixation and performance or competence factors which affect the ability to test beliefs for their truthfulness. One conclusion we can draw from the account developed here is that these epistemic approaches work best when the control hierarchy is configured for this type of epistemic task. However delusion involves a unique configuration of the control hierarchy in which the ability to test hypotheses is suppressed.

That configuration is produced by the allocation of cognitive resources whose mechanisms we explored by a comparison of the relationship between default circuitry and phenomenology and psychology in different modes: dreaming, delusion, default thought in normal waking cognition and decontextualised belief fixation. The conclusion is that delusion is a unique " mixed mode" of the control hierarchy.

Interestingly other theorists committed to versions of the autonomy thesis have reached a similar conclusion on the basis of phenomenological and psychological considerations. For example Stephens and Graham suggest that delusions are a *sui generis* metacognitive attitude produced by an "unproductive allocation of psychological resources in the

management and control of her own thinking and attitudes"(Graham and Stephens, 2007). Of course this type of misallocation is not unique to delusion. The same is true, for example of obsessive patients, who cannot prevent their obsessions from invading and monopolising memory and attention and interacting with affective and motor processes.

Consequently Stephens and Graham make another important point. Delusional subjects *identify* with their delusion. In contrast with anxious or obsessive patients, who may be distressed by their persistent thought patterns, delusional subjects are not disengaged from their delusion. Nonetheless this failure to disengage is *not* the same psychological state as being unable to distinguish veridical from non-veridical experience, as Sass among others points out.

The notion of identification is an acute observation. It seems more important to delusional subjects to stay involved with experience and the train of thought it prompts than to disengage from that experience and try and integrate it with the rest of the relevant information available to them in other cognitive contexts. They *prefer*, in the purely economic sense of resource allocation, their delusion to the truth. In contrast patients with obsessive or compulsive thought patterns may not identify with their delusion. They have a barrier to incorporation that may produce considerable distress.

Motivational accounts of delusion explain it in terms of the affective significance of the delusional belief. On these views delusions are maintained to minimise or manage distress. While it is true that delusions (and psychiatric disorders in general) have a strong emotional component what really needs to be explained is how affective states could produce a failure to dceontextualise.

A solution is suggested by the appraisal theory of emotion (Adolphs, 2003). On the appraisal theory emotions are evaluations of the *subjective relevance* of events for the subject. Automatic appraisals are coordinated by posterior circuitry centred on the amygdala. These appraisals influence higher order, explicit appraisals orchestrated by the ventromedial prefrontal cortex, *the hub of the default circuit*. This is of course not

surprising: default circuitry produces responses to salient information referred from lower levels in the control hierarchy. We might even say that what makes information essentially indexical is that it automatically evokes an affective response derived from subjective life history. This is why, as we saw, the ventromedial prefrontal cortex is needed to produce the felt subjective feeling of "mineness" which distinguishes autobiographical from purely episodic memory. It adds the felt affective tone (Gilboa, 2004).

Consequently it is not surprising that some delusions have a strong affective component. The circuitry which produces them is *intrinsically* affective because it is dedicated to the generation of autobiographical responses to experience.

Motivational accounts are in this sense too "psychological". They suggest that the delusional subject "wants" to believe her delusion. This is true in the sense that the explicit consideration of alternatives evokes distress or defensive reaction. But the cause arises earlier in the control hierarchy via the allocation of resources to processing systems whose goal is not truth but subjective adequacy.

REFERENCES

Abi-Dargham, A., Mawlawi, O., Lombardo, I., Gil, R., Martinez, D., Huang, Y., Hwang, D., Keilp, J., Kochan, L. and Van Heertum, R. 2002: Prefrontal dopamine D1 receptors and working memory in schizophrenia. ***Journal of Neuroscience,*** **22, 3708.**

Adolphs, R. 2003: Is the human amygdala specialized for processing social information? ***Annals of the New York Academy of Sciences,*** **985, 326-340.**

Adolphs, R. 2010: What does the amygdala contribute to social cognition? ***Annals of the New York Academy of Sciences,*** **1191, 42-61.**

Allen, P., Aleman, A. and Mcguire, P. K. 2007: Inner speech models of auditory verbal hallucinations: evidence from behavioural and neuroimaging studies. ***International Review of Psychiatry,*** **19, 407-415.**

Allman 1999: ***Evolving Brains,*** **New York: Scientific American Library.**

Anderson, M. 2007: The massive redeployment hypothesis and the functional topography of the brain. ***Philosophical Psychology,*** **20, 143-174.**

Andreasen, N. 1999: A unitary model of schizophrenia: Bleuler's" fragmented phrene" as schizencephaly. ***Archives of General Psychiatry,*** **56, 781.**

Andreasen, N. and Grounds, B. 1984: ***The broken brain: The biological revolution in psychiatry*****: Harper & Row New York.**

Andreasen, N. C., Oparadisio, S. and O'leary, D. S. 1998: "Cognitive Dysmetria" as an intergrative theory of schizophrenia: A dysfunction in cortical-subcortical-cerebellar circuitry? ***Schizophrenia Bulletin,*** **24, 203-218.**

Angyal, A. 1936: The experience of the body-self in schizophrenia. ***Archives of Neurology and Psychiatry,*** **35, 1029-1053.**

Arbib, M. E. 2003: ***The Handbook of Brain Theory and Neural Networks,*** **Cambridge, Mass: London: MIT Press.**

Arnsten, A. F. T. 1998: Catecholamine modulation of prefrontal cortical cognitive function. ***Trends in Cogntive Sciences,*** **2, 436-446.**

Assaf, M., Jagannathan, K., Calhoun, V., Miller, L., Stevens, M., Sahl, R., O'boyle, J., Schultz, R. and Pearlson, G. Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. ***Neuroimage*****.**

Bargh, J. and Chartrand, T. 2003: The unbearable automaticity of being. ***American Psychologist,*** **54, 462-479.**

Barto, A. G. 1994: " 1" 1 Adaptive Critics and the Basal Ganglia. ***Models of information processing in the basal ganglia*****, 215.**

Bauer, R. M. 1984: Autonomic recognition of names and faces in prosopagnosia: A neuropsychological application of the guilty knowledge test. ***Neuropsychologia,*** **22, 457-469.**

Bauer, R. M. 1986: The cognitive psychophysiology of prosopagnosia. In Ellis, H. D., Jeeves, M. A., Newcombe, F. and Young, A. (eds) ***Aspects of face Processing.*** **Dordrecht: Martinus Nijhoff.**

Bayne, T. and Pacherie, E. 2004: Bottom-Up or Top-Down? Campbell's Rationalist Account of Delusions. ***Philosophy , Psychiatry and Psychology,*** **11, 1-12.**

Bayne, T. and Pacherie, E. 2005: In defence of the doxastic conception of delusions. ***Mind & Language,*** **20, 163-188.**

Bayne, T. and Pacherie, E. 2007: Narrators and comparators: the architecture of agentive self-awareness. ***Synthese,*** **159, 475-491.**

Bechara, A., Damasio, H., Damasio, A. and Lee, G. P. 1999: Differential Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Human Decision-Making. ***The Journal of Neuroscience,*** **19, 5473-5481.**

Behrendt, R. 2004: A neuroanatomical model of passivity phenomena. ***Consciousness and cognition,*** **13, 579-609.**

Bennett, M. and Hacker, P. 2003: ***Philosophical foundations of neuroscience*****: Wiley-Blackwell.**

Bentall, R. 2010: ***Doctoring the mind*****: Allen Lane.**

Bentall, R. P. 2004: ***Madness Explained*****.**

Bentall, R. P., Kaney, S. and Bowen-Jones, K. 1995: Persecutory delusions and recall of threat-related, depression-related, and neutral words. ***Cognitive Therapy and Research,*** **19, 445-457.**

Bentall, R. P., Kaney, S. and Dewey, M. E. 1991a: Paranoia and social reasoning: An attribution theory analysis. ***British Journal of Clinical Psychology,*** **30, 13-23.**

Bentall, R. P., Kaney, S. and Dewey, M. E. 1991b: Persecutory delusions: An attribution theory analysis. ***British Journal of Clinical Psychology,*** **30, 13-23.**

Bentall, R. P., Kinderman, P. and Kaney, S. 1994: The self, attributional processes and abnormal beliefs: Towards a model of persectuory delusions. ***Behaviour Research and Therapy,*** **32, 331-341.**

Bergeron, V. 2007: Anatomical and functional modularity in cognitive science: Shifting the focus. ***Philosophical Psychology,*** **20, 175-195.**

Bermúdez, J. L. 2000a: Normativity and rationality in delusional psychiatric disorders. ***Mind and Language,*** **16, 457-493.**

Bermúdez, J. L. 2000b: Personal and subpersonal: A difference without a distinction. ***Philosophical Explorations,*** **2.**

Berridge, K. C. and Robinson, T. E. 1998: What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? ***Brain Research Reviews,*** **28, 309-369.**

Berridge, K. C. and Robinson, T. E. 2003: Parsing reward. ***Trends in Neuroscience,*** **26, 507-513.**

Berrios, G. 1992: Phenomenology, psychopathology and Jaspers: a conceptual history. ***History of Psychiatry,*** **3, 303.**

Blakemore, S., Oakley, D. and Frith, C. 2003: Delusions of alien control in the normal brain. ***Neuropsychologia,*** **41, 1058-1067.**

Blakemore, S. and Sirigu, A. 2003: Action prediction in the cerebellum and in the parietal lobe. ***Experimental Brain Research,*** **153, 239-245.**

Blakemore, S., Wolpert, D. and Frith, C. 2002: Abnormalities in the awareness of action. ***Trends in Cognitive Sciences,*** **6, 237-242.**

Blakemore, S.-J., Smith, J. and Steel, R. 2000: The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: Evidence for a breakdown in self-monitoring. ***Psychological Medicine,*** **30, 1131-1139.**

Blanke, O. and Metzinger, T. 2009: Full-body illusions and minimal phenomenal selfhood. ***Trends in Cognitive Sciences,*** **13, 7-13.**

Bortolotti, L. 2005: Delusions and the background of rationality. ***Mind & Language,*** **20, 189-208.**

Bortolotti, L. 2009: Delusion. ***Stanford Encyclopedia of Philosophy*****.**

Botzung, A., Denkova, E. and Manning, L. 2008: Experiencing past and future personal events: Functional neuroimaging evidence on the neural bases of mental time travel. ***Brain and Cognition,*** **66, 202-212.**

Braver, T. Barch, D. and Cohen, J. 1999: Cognition and Control in Schizophrenia: a Computational Model of Dopamine and Prefrontal Function. ***Biological Psychiatry,*** **46, 312-328.**

Braver, T. and Cohen, J. 1999: On the control of control: The role of dopamine in regulating prefrontal function and working memory. In Monsell, S. D., J. (ed) ***Attention and Performance XVIII.*** **Cambridge, MA: MIT Press.**

Braver, T. and Cohen, J. 2000: On the control of control: The role of dopamine in regulating prefrontal function and working memory. ***Control of cognitive processes: Attention and performance XVIII*****, 713–737.**

Breen, N., Caine, D., Coltheart, M., Hendy, J. and Roberts, C. 2000: Towards an understanding of delusions of misidentification: Four case studies. ***Mind & Language,*** **15, 74-110.**

Breen, N., Coltheart, M. and Caine, D. 2001: A two-way window on face recognition. ***Trends in Cognitive Science,*** **5, 234-235.**

Brighetti, G., Bonifacci, P., Borlimi, R. and Ottaviani, C. 2007: ìFar from the heart far from the eyeî: Evidence from the Capgras delusion. ***Cognitive neuropsychiatry,*** **12, 189-197.**

Broome, M. 2004: The Rationality of Psychosis and Understanding the Deluded ***Philosophy, Psychiatry, & Psychology,*** **11, 35-41.**

Broome, M., Woolley, J., Tabraham, P., Johns, L., Bramon, E., Murray, G., Pariante, C., Mcguire, P. and Murray, R. 2005: What causes the onset of psychosis? ***Schizophrenia research,*** **79, 23-34.**

Broome, M. R., Johns, L., Valli, I., Woolley, J., Tabraham, P., Brett, C., Valmaggia, L., Peters, E., Garety, P. and Mcguire, P. 2007: Delusion formation and reasoning biases in those at clinical high risk for psychosis. ***The British Journal of Psychiatry,*** **191, s38-s42.**

Browning, S. and Jones, S. 1988: Ichthyosis and delusions of lizard invasion. ***Acta Psychiatrica Scandinavica,*** **78, 766-767.**

Broyd, S., Demanuele, C., Debener, S., Helps, S., James, C. and Sonuga-Barke, E. 2009: Default-mode brain dysfunction in mental disorders: a systematic review. ***Neuroscience & Biobehavioral Reviews,*** **33, 279-296.**

Bruyer, R. 1991: Covert face recognition in prosopagnosia: A review. ***Brain and Cognition,*** **15, 223-235.**

Buckner, R. L., Andrews-Hanna, J. R. and Schacter, D. L. 2008: The brain's default system: Anatomy, function, and relevance to disease. ***The Year in Cognitive Neuroscience 2008, Ann. NY Acad. Sci,*** **1124, 1–38.**

Bunge, M. 2003: ***Emergence and convergence: qualitative novelty and the unity of knowledge*****: Univ of Toronto Pr.**

Butler, P. V. 2000: Diurnal variation in Cotard's syndrome (copresent with Capgras delusion) following traumatic brain injury. ***Australian and New Zealand Journal of Psychiatry,*** **34, 684-687.**

Campbell, J. 1999: Schizophrenia, the space of reasons, and thinking as a motor process. ***The Monist,*** **82, 609-625.**

Campbell, J. 2001a: Rationality Meanikng and the Analysis of Delusion. ***Philosophy, Psychiatiry adn Psychology.,*** **8, 89-100.**

Campbell, J. 2001b: Rationality, meaning and the analysis of delusion. ***Philosophy, Psychiatry and Psychology,*** **8, 89-100.**

Campbell, J. 2007: An interventionist approach to causation in psychology. ***Causal learning: Psychology, philosophy, and computation*****, 58-66.**

Carruthers, P. 2002: The cognitive functions of language. ***Behavioral and brain sciences,*** **25, 657-674.**

Chapman, L. J. and Chapman, J. P. 1988: The genesis of delusions. In Oltmans, T. F. and Maher, B. A. (eds) ***Delusional Beliefs.*** **New York: Wiley and Son.**

Clark, A. 2012: Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science.

Coltheart, M. 1999: Modularity and cognition. ***Trends in Cognitive Science,*** **3, 115-120.**

Coltheart, M. 2007: The 33rd Sir Frederick Bartlett Lecture Cognitive neuropsychiatry and delusional belief. ***The Quarterly Journal of Experimental Psychology,*** **60, 1041-1062.**

Coltheart, M. 2010: Lessons from cognitive neuropsychology for cognitive science: A reply to Patterson and Plaut (2009). ***Topics in Cognitive Science,*** **2, 3-11.**

Coltheart, M., Menzies, P. and Sutton, J. 2010: Abductive inference and delusional belief. ***Cognitive neuropsychiatry,*** **15, 261-287.**

Corlett, P., Frith, C. and Fletcher, P. 2009a: From drugs to deprivation: a Bayesian framework for understanding models of psychosis. ***Psychopharmacology,*** **206, 515-530.**

Corlett, P., Taylor, J., Wang, X. J., Fletcher, P. and Krystal, J. 2010: Toward a neurobiology of delusions. ***Progress in neurobiology,*** **92, 345-369.**

Corlett, P. R., Krystal, J. H., Taylor, J. R. and Fletcher, P. C. 2009b: Why do delusions persist? ***Frontiers in human neuroscience,*** **3.**

Cotard, J. 1882: Du délire des négation. ***Archives de Neurologie,*** **4, 152-70,282-295.**

Courbon, P. and Fail, G. 1927: Syndrome d' "illusion de Frégoli" et schizophrénie. ***Bulletin de Société Clinique de Médecine Mentale,*** **15, 121-125.**

Craver, C. 2001: Role functions, mechanisms, and hierarchy. ***Philosophy of Science,*** **68, 53-74.**

Craver, C. 2002: Interlevel experiments and multilevel mechanisms in the neuroscience of memory. ***Philosophy of Science,*** **69, 83-97.**

Craver, C. 2007: ***Explaining the brain: Mechanisms and the mosaic unity of neuroscience*****: Oxford University Press, USA.**

Csibra, G. 2005: Mirror neurons and action observation: Is simulation involved. ***What do mirror neurons mean*****.**

Currie, G. 2000a: Imagination, delusion and hallucinations. In Coltheart, M. and Davies, M. (eds) ***Pathologies of Belief.*** **Oxford: Blackwell Publishers.**

Currie, G. 2000b: Imagination, delusion and hallucinations. ***Mind & Language,*** **15, 168-183.**

Currie, G. and Jones, N. 2006: McGinn on delusion and imagination. ***Philosophical Books,*** **47, 306-313.**

Currie, G. and Ravenscroft, I. 2002: ***Recreative minds: Imagination in philosophy and psychology*****: New York.**

Currie, G. a. S., K. 2000c: How to think about the modularity of mind-reading. ***Philosophical Quarterly,*** **50, 145-60.**

D'argembeau, A., Raphael, S. and Van Der Linden, M. Disturbance of the temporal direction of self in schizophrenia. ***submitted manuscript.***

Damasio, A., Tranel, D. and Damasio, H. 1990: Individuals with sociopathic behaviour caused by frontal lobe damage fail to respond autonomically to socially charged stimuli. ***Behavioural and Brain Research,*** **14, 81-94.**

Damasio, A. R., Tranel, D. and Damasio, H. 1991: Somatic markers and the guidance of behaviour: theory and preliminary testing. In Levin, H. S., Eisenberg, H. M. and Benton (eds) ***Frontal Lobe Function and Dysfunction.*** **New York: Oxford University Press.**

Danckert, J., Rossetti, Y., D'amato, T., Dalery, J. and Saoud, M. 2002: Exploring imagined movements in patients with schizophrenia. ***Neuroreport,*** **13, 605.**

Danckert, J., Saoud, M. and Maruff, P. 2004: Attention, motor control and motor imagery in schizophrenia: implications for the role of the parietal cortex. ***Schizophrenia research,*** **70, 241-261.**

Daprati, E., Franck, N. and Georgieff, N. 1997: Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophreniz patients. ***Cognition,*** **65, 71-86.**

David, A. 2004: The cognitive neuropsychiatry of auditory verbal hallucinations: an overview. ***Voices in the Brain: The Cognitive Neuropsychiatry of Auditory Verbal Hallucinations,*** **9, 107-123.**

Davidson, D. 1980: ***Essays on Actions and Events,*** **Oxford: Oxford University Press.**

Davies, M. 1990: Thinking Persons and Cognitive Science. ***AI and Society,*** **4, 39-50.**

Davies, M. 2000a: Persons and their underpinnings. ***Philosophical Explorations,*** **3, 43-62.**

Davies, M. 2000b: Interaction without reduction: The relationship between personal and subpersonal levels of description. ***Mind and Society,*** **1, 87-105.**

Davies, M., Coltheart, M., Langdon, R. and Breen, N. 2001: Monothematic delusions: Towards a two-factor account. ***Philosophy, Psychiatry and Psychology,*** **8, 133-158.**

Dawson, J. L. and Conduit, R. 2011: The Substrate That Dreams Are Made On: An Evaluation of Current Neurobiological Theories of Dreaming. ***States of Consciousness*****, 133-156.**

De Pauw, K. W. and Szulecka, T. K. 1988: Dangerous delusions: Violence and the misidentification of syndromes. ***British Journal of Psychiatry,*** **152, 91-97.**

De Pauw, K. W., Szulecka, T. K. and Pollock, T. L. 1987: Frégoli syndrome after cerebral infarction. ***Journal of Nervous and Mental Disease,*** **175, 433-438.**

Decety, J. and Lindgren, M. 1991: Sensation of effort and duration of mentally executed actions. ***Scandinavian journal of psychology,*** **32, 97-104.**

Dennett, D. 1991: Real Patterns. ***Journal of Philosophy*****, 27-51.**

Desseilles, M., Dang-Vu, T., Sterpenich, V. and Schwartz, S. Cognitive and emotional processes during dreaming: A neuroimaging view. ***Consciousness and cognition*****.**

Devinsky, O., Davachi, L., Santchi, C., Quinn, B. T., Staresina, B. P. and Thesen, T. 2010: Hyperfamiliarity for faces. ***Neurology,*** **74, 970-974.**

Di Forti, M., Lappin, J. and Murray, R. 2007: Risk factors for schizophrenia--All roads lead to dopamine. ***European Neuropsychopharmacology,*** **17, S101-S107.**

Domhoff, G. W. 2005: Refocusing the Neurocognitive Approach to Dreams: A Critique of the Hobson Versus Solms Debate. ***Dreaming,*** **15, 3.**

Domhoff, G. W. 2011: The neural substrate for dreaming: Is it a subsystem of the default network? ***Consciousness and Cognition,*** **20, 1163-1174.**

Driver, J. 2008: Imaginative resistance and psychological necessity. ***Social Philosophy and Policy,*** **25, 301.**

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. 2006: The somatic marker hypothesis: A critical evaluation. ***Neuroscience and Biobehavioural Reviews,*** **30, 239-271.**

Durstewitz, D. and Seamans, J. 2002: The computational role of dopamine D1 receptors in working memory. ***Neural Networks,*** **15, 561-572.**

Egan, A. 2009: Imagination, delusion, and self-deception. ***Delusion and self-deception: affective and motivational influences on belief formation*****, 263.**

Egelman, D. M., Person, C., Montague, P.R. 1998: The computational role of dopamine delivery. . ***Journal of Cognitive Neuroscience.,*** **10, 623-630.**

Eidelberg, D. and Galaburda, A. M. 1984: Inferior parietal lobule: Divergent architectonic asymmetries in the human brain. ***Archives of Neurology,*** **41, 843-852.**

Ellis, H. D. and De Pauw, K. W. 1994: The cognitive neuropsychiatric origins of the Capgras delusion. In David, A. and Cutting, J. (eds) ***Cognitive Neuropsychology of Schizophrenia.*** **Hove, UK: Lawrence Erlbaum.**

Ellis, H. D., De Pauw, K. W., Christodoulou, G. N., Papageorgiou, L., Milne, A. B. and Joseph, A. B. 1993a: Responses to facial and non-facial stimuli presented tachistoscopically in either or both visual fields by patients with the Capgras delusion and paranoid schizophrenics. ***Journal of Neurology, Neurosurgery, and Psychiatry,*** **56, 215-219.**

Ellis, H. D. and Lewis, M. B. 2001: Capgras delusions: A window on face recognition. ***Trends in Cognitive Science,*** **5, 149-156.**

Ellis, H. D. and Szulecka, T. K. 1996: The disguised lover: A case of Frégoli delusion. In Halligan, P. W. and Marshall, J. C. (eds) ***Method in Madness: Case studies in cognitive neuropsychiatry.*** **Hove, East Sussex: Psychology Press.**

Ellis, H. D., Whitley, J. and J.-P., L. 1994: Delusional misidentifications:The three original papers on the Capgras, Frégoli and Intermetamorphosis delusions. ***History of Psychiatry,*** **5, 117-146.**

Ellis, H. D., Young, A. W. and Koenken, G. 1993b: Covert face recognition without prosopagnosia. ***Behavioural Neurology,*** **6, 27-32.**

Enoch, M. D. and Trethowan, W. H. 1991: ***Uncommon Psychiatric Syndromes,*** **Oxford: Butterworth-Heinemann.**

Eva, F. J. and Perry, D. 1993: The Frégoli syndrome and cannabis delusional disorder. ***British Journal of Psychological Medicine,*** **10, 87-88.**

Fair, D., Cohen, A., Dosenbach, N., Church, J., Miezin, F., Barch, D., Raichle, M., Petersen, S. and Schlaggar, B. 2008: The maturing architecture of the brain's default network. ***Proceedings of the National Academy of Sciences,*** **105, 4028.**

Fodor, J. 1983: ***The Modularity of Mind,*** **Cambridge, MA: MIT Press.**

Fodor, J. and Lepore, E. 1992: ***Holism: A shopper's guide*****: Wiley-Blackwell.**

Ford, J. and Mathalon, D. 2004: Electrophysiological evidence of corollary discharge dysfunction in schizophrenia during talking and thinking. ***Journal of Psychiatric Research,*** **38, 37-46.**

Ford, J. and Mathalon, D. 2005: Corollary discharge dysfunction in schizophrenia: can it explain auditory hallucinations? ***International Journal of Psychophysiology,*** **58, 179-189.**

Ford, J., Roach, B., Faustman, W. and Mathalon, D. 2007: Synch before you speak: auditory hallucinations in schizophrenia. ***American Journal of Psychiatry,*** **164, 458.**

Friston, K. 2003: Learning and inference in the brain. ***Neural Networks,*** **16, 1325-1352.**

Frith, C., Blakemore, S. and Wolpert, D. 2000: Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. ***Brain Research Reviews,*** **31, 357-363.**

Fuster, J. 2008: ***The prefrontal cortex*****: Academic Pr.**

Gallagher, S. 2004: Neurocognitive models of schizophrenia: a neurophenomenological critique. ***Psychopathology,*** **37, 8-19.**

Gallese, V., Fadiga, L., Fogassi, L. and Rizzolatti, G. 1996: Action recognition in the premotor cortex. ***Brain,*** **119 (Pt 2), 593-609.**

Gallese, V. and Goldman, A. 1998: Mirror neurons and the simulation theory of mind-reading. ***Trends in Cognitive Sciences,*** **2, 493-501.**

Gallese, V., Keysers, C. and Rizzolatti, G. 2004: A unifying view of the basis of social cognition. ***Trends Cogn Sci,*** **8, 396-403.**

Ganesan, V., Hunter, M. and Spence, S. 2005: Schneiderian first-rank symptoms and right parietal hyperactivation: A replication using FMRI. ***American Journal of Psychiatry,*** **162, 1545.**

Garety, P. A. and Freeman, D. 1999: Cognitive approaches to delusions: A critical review of theories and evidence. ***British Journal of Clinical Psychology,*** **38, 113-154.**

Garety, P. A., Hemsley, D. R. and Wessely, S. 1991: Reasoning in deluded schizophrenic and paranoid patients: Biases in performance on a probabilistic inference task. ***Journal of Nervous and Mental Disease,*** **179, 194-201.**

Gendler, T. 2011: Imagination.

Gendler, T. S. 2000: The puzzle of imaginative resistance. ***Journal of Philosophy,*** **97, 55-81.**

Gendler, T. S. 2008a: Alief and belief. ***Journal of Philosophy,*** **105, 634-663.**

Gendler, T. S. 2008b: Alief in action (and reaction). ***Mind & Language,*** **23, 552-585.**

Georgieff, N. and Jeannerod, M. 1998: Beyond consciousness of external reality: A "Who" system for consciousness of action and self-consciousness. ***Consciousness and Cognition,*** **7, 465-477.**

Gerrans, P. 1999: Delusional misidentification as subpersonal disintegration. ***The Monist,*** **82, 590-608.**

Gerrans, P. 2000: Refining the explanation of Cotard's delusion. ***Mind and Language,*** **15, 111-122.**

Gerrans, P. 2001: Delusions as performance failures. ***Cognitive Neuropsychiatry,*** **6, 161-173.**

Gerrans, P. 2002a: Nativism and Neuroconstructivism in the explanation of Williams Syndrome. ***Biology and Philosophy,*** **17, 1-11.**

Gerrans, P. 2002b: A One-Stage Theory of The Cotard Delusion. 9, 47-53.

Gerrans, P. 2003: A one-stage explanation of the Cotard delusion. ***Philosophy, Psychiatry, & Psychology,*** **9, 47-53.**

Gerrans, P. 2010a: Imitation and Theory of Mind. In Berntson, G. T. and Cacioppo, J. G. (eds) ***Handbook of neuroscience for the behavioral sciences, Volume 2.*** **Chicago: Wiley.**

Gerrans, P. 2010b: Mirror neuron systems: The role of mirroring processes in social cognition. ***Cognitive neuropsychiatry,*** **15, 501-504.**

Gerrans, P. 2012: Experience and Expectations: Bayesian Explanations of the Alternation Between the Capgras and Cotard Delusions. ***Philosophy, Psychiatry, & Psychology,*** **19, 145-148.**

Gilbert, D. 2004: The Big Wombassa.

Gilbert, S., Frith, C. and Burgess, P. 2005: Involvement of rostral prefrontal cortex in selection between stimulus oriented and stimulus independent thought. ***European Journal of Neuroscience,*** **21, 1423-1431.**

Gilboa, A. 2004: Autobiographical and episodic memory,Äîone and the same?: Evidence from prefrontal activation in neuroimaging studies. *Neuropsychologia,* 42, 1336-1349.

Gilboa, A., Winocur, G., Grady, C., Hevenor, S. and Moscovitch, M. 2004: Remembering our past: functional neuroanatomy of recollection of recent and very remote personal events. *Cerebral Cortex,* 14, 1214-1225.

Gilleen, J. D., A. 2004: The cognitive neuropsychiatry of delusions: from psychopathology to neuropsychology and back again. *Psychological Medicine,* 35, 5-12.

Goel, V. 2004: Logical reasoning deficits in schizophrenia. *Schizphrenia Research,* 66, 87-88.

Goel, V. 2007: Anatomy of deductive reasoning. *Trends in Cognitive Sciences,* 11, 435-441.

Goel, V. D., R. 2000: Anatomical Segregation of Component Processes in an Inductive Inference Task. *Journal of Cognitive Neuroscience,* 12, 110-119.

Goldman, A. 2006: *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*: Oxford University Press, USA.

Goldman-Rakic, P. 1997: The cortical dopamine system: role in memory and cognition. *Advances in pharmacology,* 42, 707-711.

Gottesmann, C. 2006: The dreaming sleep stage: A new neurobiological model of schizophrenia? *Neuroscience,* 140, 1105-1115.

Gottfried, J., O'doherty, J. and Dolan, R. 2003: Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science,* 301, 1104.

Grace, A. 1991: Phasic versus tonic dopamine release and the modulation of dopamine responsivity: A hypothesis of the etiology of schizophrenia. *Neuroscience,* 41, 1-24.

Graham, G. 2009: *The disordered mind: an introduction to philosophy of mind and mental illness*: Routledge.

Graham, G. and Stephens, L. 2007: The Delusional Stance. Reconceiving Schizophrenia, New York: Oxford University Press.

Grézes, J. and Decety, J. 2001: Functional anatomy of execution, mental simulation, observation and verb generation of actions: A meta-analysis. *Human Brain Mapping,* 12, 1-19.

Gusnard, D., Akbudak, E., Shulman, G. and Raichle, M. 2001: Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences,* 98, 4259.

Halligan, P. W. and Marshall, J. C. (eds) 1996: *Method in Madness: Case studies in cognitive neuropsychiatry.,* Hove, East Sussex: Psychology Press.

Hartmann, E., Milofsky, E. and Vaillant, G. 1984: Vulnerability to schizophrenia: Prediction of adult schizophrenia using childhood information. *Archives of General Psychiatry,* 41, 1050-1056.

Hassabis, D., Kumaran, D. and Maguire, E. A. 2007a: Using imagination to understand the neural basis of episodic memory. *The Journal of Neuroscience,* 27, 14365-14374.

Hassabis, D., Kumaran, D., Vann, S. D. and Maguire, E. A. 2007b: Patients with hippocampal amnesia cannot imagine new experiences. ***Proceedings of the National Academy of Sciences,*** **104, 1726.**

Hassabis, D. and Maguire, E. A. 2007a: Deconstructing episodic memory with construction. ***Trends in cognitive sciences,*** **11, 299-306.**

Hauser, M. D. 2006: The liver and the moral organ. ***Social Cognitive and Affective Neuroscience,*** **1, 214-220.**

Heinks-Maldonado, T., Mathalon, D., Houde, J., Gray, M., Faustman, W. and Ford, J. 2007: Relationship of imprecise corollary discharge in schizophrenia to auditory hallucinations. ***Archives of General Psychiatry,*** **64, 286.**

Heinz, A. and Schlagenhauf, F. 2010: Dopaminergic dysfunction in schizophrenia: salience attribution revisited. ***Schizophrenia bulletin,*** **36, 472-485.**

Hemsley, D. R. and Garety, P. A. 1986: The formation and maintenance of delusions: A Bayesian analysis. ***British Journal of Psychiatry,*** **149, 51-56.**

Hobson, J. 1999: ***Dreaming as delirium: How the brain goes out of its mind*****: The MIT Press.**

Hohwy, J. 2004: Top-Down and Bottom-Up in Delusion Formation. ***Philosophy , Psychiatry and Psychology,*** **11, 65-70.**

Hohwy, J., Roepstorff, A. and Friston, K. 2008: Predictive coding explains binocular rivalry: An epistemological review. ***Cognition,*** **108, 687-701.**

Horovitz, S., Braun, A., Carr, W., Picchioni, D., Balkin, T., Fukunaga, M. and Duyn, J. 2009: Decoupling of the brain's default mode network during deep sleep. ***Proceedings of the National Academy of Sciences,*** **106, 11376.**

Howes, O. and Kapur, S. 2009: The dopamine hypothesis of schizophrenia: version III--the final common pathway. ***Schizophrenia Bulletin*****.**

Huq, S. F., Garety, P. A. and Hemsley, D. R. 1988: Probabilistic judgements in deluded and non-deluded subjects. ***Quarterly Journal of Experimental Psychology,*** **40, 801-813.**

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C. and Rizzolatti, G. 2005: Grasping the intentions of others with one's own mirror neuron system. ***PLoS Biol,*** **3, e79.**

Ichikawa, J. 2008: Scepticism and the imagination model of dreaming. ***The Philosophical Quarterly,*** **58, 519-527.**

Ichikawa, J. 2009: Dreaming and imagination. ***Mind & Language,*** **24, 103-121.**

Jackson, H. 1958: Selected writings of Hughlings Jackson (Vol. 2). New York: Basic Books.(Original work published 1866).

Jaspers, K. 1963: ***General Psychopathology,*** **Manchester: Manchester University Press.**

Jeannerod, M. 2006: ***Motor cognition: What actions tell the self*****: Oxford University Press, USA.**

Jeannerod, M., Farrer, C., Franck, N., Fourneret, P., Posada, A., Daprati, E. and Georgieff, N. 2003: Action recognition in normal and schizophrenic subjects. ***The self in neuroscience and psychiatry*****, 380ñ406.**

Jeannerod, M. and Frak, V. 1999: Mental Imaging of Motor Activity in Humans. ***Current Opinion in Neurobiology,*** **9, 735-739.**

Jones, M. and Love, B. C. 2011: Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. ***Behavioral and Brain Sciences,*** **34, 169-188.**

Jones, S. R. 2010: Do we need multiple models of auditory verbal hallucinations? Examining the phenomenological fit of cognitive and neurological models. ***Schizophrenia bulletin,*** **36, 566-575.**

Jones, S. R. and Fernyhough, C. 2007a: Neural correlates of inner speech and auditory verbal hallucinations: a critical review and theoretical integration. ***Clinical psychology review,*** **27, 140-154.**

Jones, S. R. and Fernyhough, C. 2007b: Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. ***Consciousness and Cognition,*** **16, 391-399.**

Joseph, A. B. and O'leary, D. H. 1987: Anterior cortical atrophy in Frégoli syndrome. ***Journal of Clinical Psychiatry,*** **48, 409-411.**

Kahn, D., Pace-Schott, E. and Hobson, J. A. 2002: Emotion and cognition: Feeling and character identification in dreaming. ***Consciousness and Cognition,*** **11, 34-50.**

Kaney, S. and Bentall, R. P. 1989: Persecutory delusions and attributional style. ***British journal of Medical Psychology,*** **62, 191-198.**

Kaney, S. and Bentall, R. P. 1992: Persecutory delusions and the self-serving bias: Evidence from a contigency judgement task. ***Journal of Nervous and Mental Disease,*** **180, 773-780.**

Kaney, S., Wolfenden, M., Dewey, M. E. and Bentall, R. P. 1992: Persecutory delusions and recall of threatening and non-threatening propositions. ***British Journal of Clinical Psychology,*** **31, 85-87.**

Kapur, J. 2003: Psychosis as a State of Aberrant Salience: A Framework for Linking Biology, Phenomenology and Pharmacology in Sschizophrenia. ***American Journal of Psychiatry,*** **160, 13-23.**

Kapur, S. 2004: How antipsychotics become anti-[] psychotic'-from dopamine to salience to psychosis. ***Trends in pharmacological sciences,*** **25, 402-406.**

Kapur, S., Mizrahi, R. and Li, M. 2005: From dopamine to salience to psychosis--linking biology, pharmacology and phenomenology of psychosis. ***Schizophrenia research,*** **79, 59-68.**

Karmiloff-Smith, A. 1994: Beyond modularity: A developmental perspective on cognitive science. ***International Journal of Language & Communication Disorders,*** **29, 95-105.**

Kennedy, D. and Courchesne, E. 2008: Functional abnormalities of the default network during self-and other-reflection in autism. ***Social cognitive and affective neuroscience,*** **3, 177.**

Kinderman, P. and Bentall, R. P. 1996: A new measure of causal locus: The internal, personal, and situational attributions questionnaire. ***Personality and Individual Differences,*** **20, 261-264.**

Kinderman, P. and Bentall, R. P. 1997: Causal attributions in paranoia and depression: Internal, personal and situational attributions for negative events. ***Journal of Abnormal Psychology,*** **106, 341-345.**

Kinderman, P., Kaney, S., Morley, S. and Bentall, R. P. 1992: Paranoia and the defensive attributional style: Deluded and depressed patients' attributions about their own attributions. ***British journal of Medical Psychology,*** **65, 371-383.**

Klee, R. 2004: Why some delusions are necessarily inexplicable beliefs. ***Philosophy, Psychiatry, & Psychology,*** **11, 25-34.**

Knight, R. T. 1999: Prefrontal cortex regulates inhibition and excitation in distributed neural networrks. ***Acta Psychiatrica Scandinavia,*** **101, 159-178.**

Koenigs, M. and Grafman, J. 2009: The functional neuroanatomy of depression: distinct roles for ventromedial and dorsolateral prefrontal cortex. ***Behavioural brain research,*** **201, 239-243.**

Kraepelin, E. 1896: ***Psychiatrie,*** **Leipzig: J.A. Barth.**

Lafargue, G., D'amico, A., Thobois, S., Broussolle, E. and Sirigu, A. 2008: The ability to assess muscular force in asymmetrical Parkinson's disease. ***Cortex; a journal devoted to the study of the nervous system and behavior,*** **44, 82.**

Lafargue, G. and Franck, N. 2009: Effort awareness and sense of volition in schizophrenia. ***Consciousness and cognition,*** **18, 277-289.**

Lafargue, G., Paillard, J., Lamarre, Y. and Sirigu, A. 2003: Production and perception of grip force without proprioception: is there a sense of effort in deafferented subjects? ***European Journal of Neuroscience,*** **17, 2741-2749.**

Lafargue, G. and Sirigu, A. 2006: The nature of the sense of effort and its neural substratum. ***Revue neurologique,*** **162, 703-712.**

Landis, T., Cummings, J. L., Christen, L., Bogen, J. E. and Imhof, H.-G. 1986: Are unilateral right posterior cerebral lesions sufficent to cause prosopagnosia? Clinical and radiological findings in six additional patients. ***Cortex,*** **22, 243-252.**

Langland-Hassan, P. 2008: Fractured phenomenologies: thought insertion, inner speech, and the puzzle of extraneity. ***Mind & Language,*** **23, 369-401.**

Lenggenhager, B., Tadi, T., Metzinger, T. and Blanke, O. 2007: Video ergo sum: manipulating bodily self-consciousness. ***Science,*** **317, 1096.**

Lestou, V., Pollick, F. E. and Kourtzi, Z. 2008: Neural substrates for action understanding at different description levels in the human brain. ***Journal of Cognitive Neuroscience,*** **20, 324-341.**

Levy, N. 2005: Imaginative resistance and the moral/conventional distinction. ***Philosophical Psychology,*** **18, 231-241.**

Lewis, D. 1980: Mad pain and Martian pain. ***Readings in the Philosophy of Psychology,*** **1, 216-222.**

Lieberman, M., Gaunt, R., Gilbert, D. and Trope, Y. 2002: Reflexion and reflection: A social cognitive neuroscience approach to attributional inference. ***Advances in experimental social psychology,*** **34, 199-249.**

Lieberman, M., Jarcho, J. and Satpute, A. 2004: Evidence-based and intuition-based self-knowledge: An fMRI study. ***Journal of Personality and Social Psychology,*** **87, 421-435.**

Lupien, S., Fiocco, A., Wan, N., Maheu, F., Lord, C., Schramek, T. and Tu, M. 2005: Stress hormones and human memory function across the lifespan. ***Psychoneuroendocrinology,*** **30, 225-242.**

Lupien, S., Maheu, F., Tu, M., Fiocco, A. and Schramek, T. 2007: The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. ***Brain and cognition,*** **65, 209-237.**

Lyon, H. M., Kaney, S. and Bentall, R. P. 1994: The defensive function of persecutory delusions: Evidence from attribution tasks. ***British Journal of Psychiatry,*** **164, 637-646.**

Macdonald, P. and Paus, T. 2003: The role of parietal cortex in awareness of self-generated movements: a transcranial magnetic stimulation study. ***Cerebral Cortex,*** **13, 962.**

Maguire, E. A. and Hassabis, D. 2011: Role of the hippocampus in imagination and future thinking. ***Proceedings of the National Academy of Sciences,*** **108, E39.**

Maguire, E. A., Vargha-Khadem, F. and Hassabis, D. 2010: Imagining fictitious and future experiences: Evidence from developmental amnesia. ***Neuropsychologia,*** **48, 3187-3192.**

Maher, B. A. 1988: Anamolous experiences and delusional thinking: The logic of explanation. In Oltmans, T. H. and Maher, B. A. (eds) ***Delusional Beliefs: Interdisciplinary Perspectives.*** **New York: Wiley-Interscience.**

Maher, B. A. 1999: Anomalous experience in everday life: Its significance for psychopathology. ***The Monist,*** **82, 547-570.**

Maia, V. a. C., A. 2005: Consciousness: converging insights from connectionist modeling and neuroscience. ***Trends in Cognitive Sciences,*** **9, 397-404.**

Malloy, P. F. and Richardson, E. D. 1994: The frontal lobes and content-specific delusions. ***The Journal of Neuropsychiatry and Clinical Neurosciences; The Journal of Neuropsychiatry and Clinical Neurosciences*****.**

Marr, D. 1982: ***Vision,*** **Boston: Houghtnon Mifflin.**

Maruff, P. and Currie, J. in pres: Abnormalities of motor imagery associated with somatic passivity phenomena in schizophrenia. ***Schizophrenia Research*****.**

Maruff, P., Wilson, P. and Currie, J. 2003: Abnormalities of motor imagery associated with somatic passivity phenomena in schizophrenia. ***Schizophrenia research,*** **60, 229-238.**

Mason, M., Norton, M., Van Horn, J., Wegner, D., Grafton, S. and Macrae, C. 2007: Wandering minds: the default network and stimulus-independent thought. ***Science,*** **315, 393.**

Mcclelland, J. L., Mcnaughton, B. L. and O'reilly, R. C. 1995: Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. ***Psychological review,*** **102, 419.**

Mcclure, S. M., Daw, N. D. and Read Montague, P. 2003: A computational substrate for incentive salience. ***Trends in Neurosciences,*** **26, 423-428.**

Mcdowell, J. H. 1996: ***Mind and world: with a new Introduction*****: Harvard Univ Pr.**

Mcginn, C. 2004: ***Mindsight: Image, dream, meaning*****: Harvard Univ Pr.**

Mckay, R. and Cipolotti, L. 2007: Attributional style in a case of Cotard delusion. ***Consciousness and cognition,*** **16, 349-359.**

Mercier, H. and Sperber, D. 2011a: Argumentation: Its adaptiveness and efficacy. ***Behavioral and Brain Sciences,*** **34, 94-111.**

Mercier, H. and Sperber, D. 2011b: Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences,* 34, 57.

Miller, E. and Cohen, J. D. 2001: An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience,* 24, 167-202.

Mitchell, K. and Johnson, M. 2000: Source monitoring: Attributing mental experiences. *The Oxford handbook of memory*, 179ñ195.

Mlakar, J., Jensterle, K. and Frith, C. 1994: Central monitoring deficency and schizophenic symptoms. *Psychological Medicine,* 24, 557-564.

Montague, P. R., Dayan, P. and Sejnowski, T. J. 1996: A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience,* 16, 1936-1947.

Moore, H. and Grace, A. 1999: The Regulation of Forebrain Dopamine Transmission: Relevance to the Pathophysiology and Psychopathology of Schizophrenia. *Biological Psychiatry,* 46, 40-55.

Moritz, S., Woodward, T. and Ruff, C. 2003: Source monitoring and memory confidence in schizophrenia. *Psychological medicine,* 33, 131-139.

Mujica-Parodi, L. R., Malapisina, D., Sackheim, H.A. 2000: Logical processing, affect and delusioon in schizophrenia. *Harvard Review of Psychiatry,* 8, 73-83.

Munakata, Y. and Mcclelland, J. L. 2003: Connectionist models of development. *Developmental Science,* 6, 413-429.

Murphy, D. 2006: *Psychiatry in the scientific image*: MIT Press.

Nichols, S. 2004: Imagining and believing: The promise of a single code. *The Journal of aesthetics and art criticism,* 62, 129-139.

Nichols, S. 2006: *The architecture of the imagination: New essays on pretence, possibility, and fiction*: Oxford University Press, USA.

Nichols, S. 2008: Imagination and the I. *Mind & Language,* 23, 518-535.

O'brien, G. and Opie, J. 1999: A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences,* 22, 127-148.

O'connor, A. R. and Moulin, C. J. A. 2010: Recognition without identification, erroneous familiarity, and d√ ©j√ † vu. *Current psychiatry reports,* 12, 165-173.

O'reilly, R. and Munakata, Y. 2000: *Computational Explorations in Cognitive Neurosciecne,* Cambridge, Mass:: MIT Press.

Oztop, E., Kawato, M. and Arbib, M. 2006: Mirror neurons and imitation: a computationally guided review. *Neural Netw,* 19, 254-71.

Pacherie, E. 2000: The content of intentions. *Mind and Language,* 15, 400-432.

Pacherie, E. 2001: Agency lost and found: A commentary on Spence. *Philosophy, Psychiatry and Psychology,* 8, 173-176.

Pacherie, E., Green, M. and Bayne, T. 2006: Phenomenology and delusions: Who put the "alien" in alien control? *Consciousness and Cognition,* 15, 566-577.

Pankow, A., Knobel, A., Voss, M. and Heinz, A. 2012: Neurobiological Correlates of Delusion: Beyond the Salience Attribution Hypothesis. *Neuropsychobiology,* 66, 33-43.

Papageorgiou, C., Lykouras, L., Alevizos, B., Ventouras, E., Mourtzouchou, P., Uzunoglu, N., Christodoulou, G. and Rabavilas, A. 2005: Psychophysiological differences in schizophrenics with and without delusional misidentification

syndromes: a P300 study. *Progress in Neuropsychopharmacology & Biological Psychiatry,* 29, 593-601.

Papageorgiou, C., Ventouras, E., Lykouras, L., Uzunoglu, N. and Christodoulou, G. 2003: Psychophysiological evidence for altered information processing in delusional misidentification syndromes. *Progress in Neuro-Psychopharmacology and Biological Psychiatry,* 27, 365-372.

Parnas, J. and Sass, L. 2002: Phenomenology of self-disturbances in schizophrenia: Some research findings and directions. *Philosophy, Psychiatry, & Psychology,* 8, 347-356.

Parnas, J., Sass, L. and Zahavi, D. 2008: Recent developments in philosophy of psychopathology. *Current Opinion in Psychiatry,* 21, 578.

Parnas, J. and Sass, L. A. 2001: Self, solipsism, and schizophrenic delusions. *Philosophy, Psychiatry and Psychology,* 8, 101-120.

Patterson, K. and Plaut, D. C. 2009: ,ÄúShallow draughts intoxicate the brain,Äù: Lessons from Cognitive Science for Cognitive Neuropsychology. *Topics in Cognitive Science,* 1, 39-58.

Plaut, D. 1995: Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology,* 17, 291-321.

Plaut, D. C. and Patterson, K. 2010: Beyond functional architecture in cognitive neuropsychology: A reply to Coltheart (2010). *Topics in Cognitive Science,* 2, 12-14.

Pomarol-Clotet, E., Salvador, R., Sarro, S., Gomar, J., Vila, F., Martinez, A., Guerrero, A., Ortiz-Gil, J., Sans-Sansa, B. and Capdevila, A. 2008: Failure to deactivate in the prefrontal cortex in schizophrenia: dysfunction of the default mode network? *Psychological medicine,* 38, 1185-1193.

Proust, J. 2006: Rationality and metacognition in non-human animals.

Proust, J. 2007: Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? *Synthese,* 159, 271-295.

Raichle, M. and Snyder, A. 2007: A default mode of brain function: a brief history of an evolving idea. *Neuroimage,* 37, 1083-1090.

Ramachandran, V. and Rogers-Ramachandran, D. 1996: Synaesthesia in phantom limbs induced with mirrors. *Proceedings: Biological Sciences,* 263, 377-386.

Ramachandran, V. S. and Hirstein, W. 1998: The perception of phantom limbs. The D.O. Hebb lecture. *Brain,* 121, 1603-1630.

Ratcliffe, M. Delusional Atmosphere and Delusional Belief. *Handbook of Phenomenology and Cognitive Science*, 575-590.

Ratcliffe, M. 2004: Interpreting delusions. *Phenomenology and the cognitive sciences,* 3, 25-48.

Revonsuo, A. 2006: Inner presence: Consciousness as a biological phenomenon.

Revonsuo, A. and Salmivalli, C. 1995: A content analysis of bizarre elements in dreams. *Dreaming*.

Revonsuo, A. and Tarkko, K. 2002: Binding in Dreams-The Bizarreness of Dream Images and the Unity of Consciousness. *Journal of Consciousness Studies,* 9, 3-24.

Rizzolatti, G., Ferrari, P. F., Rozzi, S. and Fogassi, L. 2006: The inferior parietal lobule: where action becomes perception. ***Novartis Found Symp,*** **270, 129-40; discussion 140-5, 164-9.**

Roehrenbach, C. and Landis, T. 1995: Dreamjourneys: Living in woven realities, the syndrome of reduplicative paramnesia. In Campbell, R. and Conway, M. A. (eds) ***Broken Memories: Case studies in memory impairment.*** **Oxford: Blackwell.**

Rumelhart, D. E. 1998: The architecture of mind: A connectionist approach. ***Mind readings*****, 207-238.**

Rushworth, M. and Taylor, P. 2006: TMS in the parietal cortex: updating representations for attention and action. ***Neuropsychologia,*** **44, 2700-2716.**

Sajonz, B., Kahnt, T., Margulies, D. S., Park, S. Q., Wittmann, A., Stoy, M., Str√ ∂ Hle, A., Heinz, A., Northoff, G. and Bermpohl, F. 2010: Delineating self-referential processing from episodic memory retrieval: Common and dissociable networks. ***Neuroimage,*** **50, 1606-1617.**

Saoud, M., Coello, Y., Dumas, P., Franck, N., D Amato, T., Dalery, J. and Rossetti, Y. 2000: Visual pointing and speed/accuracy trade-off in schizophrenia. ***Cognitive neuropsychiatry,*** **5, 123-134.**

Sass, L. 1992: Madness and modernism: Insanity in the light of modern art, literature, and thought. New York: Basic Books.

Sass, L. 1994: ***The paradoxes of delusion: Wittgenstein, Schreber, and the schizophrenic mind*****: Cornell Univ Pr.**

Sass, L. and Parnas, J. 2003: Schizophrenia, consciousness, and the self. ***Schizophrenia Bulletin,*** **29, 427.**

Schacter, D. L., Addis, D. R. and Buckner, R. L. 2007: Remembering the past to imagine the future: the prospective brain. ***Nature Reviews Neuroscience,*** **8, 657-661.**

Schacter, D. L., Addis, D. R. and Buckner, R. L. 2008: Episodic simulation of future events: concepts, data, and applications. ***Annals of the New York Academy of Sciences,*** **1124, 39-60.**

Schultz, W., Dayan, P & Montague, Pr 1997: A neural substrate of prediction and reward. Science 275:1593-1599. ***Science,*** **275, 1593-1599.**

Schwabe, L. and Blanke, O. 2007: Cognitive neuroscience of ownership and agency. ***Conscious Cogn,*** **16, 661-666.**

Schwartz, S. and Maquet, P. 2002: Sleep imaging and the neuro-psychological assessment of dreams. ***Trends in Cognitive Sciences,*** **6, 23-30.**

Schweinberger, S. and Burton, M. 2003: Covert Recognition and the Neural System for Face Processing. ***Cortex,*** **39, 9-30.**

Seamans, J. and Yang, C. 2004: The principal features and mechanisms of dopamine modulation in the prefrontal cortex. ***Progress in Neurobiology,*** **74, 1-58.**

Searle, J. R. 1983: ***Intentionality: An essay in the philosophy of mind,*** **Cambridge, UK: Cambridge University Press.**

Seligman, M. E. P., Abramson, L. Y., Semmel, A. and Von Baeyer, C. 1979: Depressive attributional style. ***Journal of Abnormal Psychology,*** **88, 242-247.**

Sergent, J. and Signoret, J.-L. 1992: Functional and anatomical decomposition of face processing: Evidence from prosopagnosia and PET study of normal subjects. *Philosophical Transactions of the Royal Society, London,* B335.

Shallice, T. 1988: *From Neuropsychology to Mental Structure,* Cambridge: Cambridge University Press.

Shergill, S., Brammer, M., Fukuda, R., Williams, S., Murray, R. and Mcguire, P. 2003: Engagement of brain areas implicated in processing inner speech in people with auditory hallucinations. *The British Journal of Psychiatry,* 182, 525.

Shraberg, D. and Weitzel, W. D. 1979: Prosopagnosia and the Capgras syndrome. *Journal of Clinical Psychiatry,* 40, 313-316.

Singh, S., Barto, A. G. and Chentanez, N. 2005: Intrinsically motivated reinforcement learning. DTIC Document.

Sirigu, A., Daprati, E., Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A. and Haggard, P. 2003: Altered awareness of voluntary action after damage to the parietal cortex. *Nature Neuroscience,* 7, 80-84.

Smith, A., Li, M., Becker, S. And Kapur, S. 2006: Dopamine, prediction error, and associative learning: a model-based account. . *Network: Computation in Neural Systems* 17, 61-84.

Sno, H. N. 2000: Deja vu and jamais vu. *Memory disorders in psychiatric practice*, 338-347.

Solms, M. 2000: Dreaming and REM sleep are controlled by different brain mechanisms. *Behavioral and brain sciences,* 23, 843-850.

Solms, M. 2007: The Interpretation of Dreams and the Neurosciences. *Developmental science and psychoanalysis: integration and innovation*, 141.

Spence, S. 2002: Alien motor phenomena: A window on to agency. *Cognitive neuropsychiatry,* 7, 211-220.

Spence, S., Brooks, D., Hirsch, S., Liddle, P., Meehan, J. and Grasby, P. 1997a: A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control). *Brain,* 120.

Spence, S. A., Brooks, D. J., Hirsch, S. R., Liddle, P. F., Meehan, J. and Grasby, P. M. 1997b: A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena. *Brain,* 120, 1997-2013.

Spier, S. 1992: Capgras syndrome and the delusions of misidentification. *Psychiatric Annals,* 22, 279-285.

Spreng, R., Mar, R. and Kim, A. 2009: The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of Cognitive Neuroscience,* 21, 489-510.

Startup, H., Freeman, D. and Garety, P. A. 2008: Jumping to conclusions and persecutory delusions. *European Psychiatry,* 23, 457-459.

Stein, D. J. and Ludik, J. 1998: *Neural networks and psychopathology: connectionist models in practice and research*: Cambridge Univ Pr.

Stein, E. 1996: *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science,* Oxford: Oxford University Press.

Stone, M. and Young, A. W. 1997: Delusions and brain injury: The philosophy and psychology of belief. ***Mind and Language,*** **13, 327-364.**

Suddendorf, T. 2006: Foresight and evolution of the human mind.

. ***Science,*** **312, 1006-1007.**

Sutton, J. 2004: Representation, reduction, and interdisciplinarity in the sciences of memory. ***Representation in mind: New approaches to mental representation*****, 187-216.**

Synofzik, M., Thier, P., Leube, D., Schlotterbeck, P. and Lindner, A. 2009: Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. ***Brain*****.**

Synofzik, M., Vosgerau, G. and Newen, A. 2008: I move, therefore I am: A new theoretical framework to investigate agency and ownership. ***Consciousness and cognition,*** **17, 411-424.**

Thompson, E., Palacios, A. and Varela, F. 2002: Ways of coloring: Comparative color vision as a case study in cognitive science. ***Vision and Mind: Selected Readings in the Philosophy of Perception, The MIT Press, Cambridge (MA)*****.**

Thornton, T. T. 2000: Mental illness and reductionism: Can functions be naturalized? ***Philosophy, Psychiatry, & Psychology,*** **7, 67-76.**

Todd, C. S. 2009: Imaginability, morality, and fictional truth: dissolving the puzzle of ‚Äòimaginative resistance‚Äô. ***Philosophical studies,*** **143, 187-211.**

Vaillancourt, D., Thulborn, K. and Corcos, D. 2003: Neural basis for the processes that underlie visually guided and internally guided force control in humans. ***Journal of Neurophysiology,*** **90, 3330.**

Vygotsky, L. 1986: Thought and language (A. Kozulin, trans.). Cambridge, Ma: MIT Press.

Vygotsky, L. S. 1978: Mind and society: The development of higher mental processes. Cambridge, MA: Harvard University Press.

Waelti, P., Dickinson, A. and Schultz, W. 2001: Dopamine responses comply with basic assumptions of formal learning theory. ***Nature,*** **412, 43-48.**

Walton, K. 2006: On the (so-called) puzzle of imaginative resistance. ***S. Nichols (Ed.)*****, 137-148.**

Waltz, J., Knowlton, B., Holyoak, K., Boone, K., Mishkin, F., De Menezes Santos, M., Thomas, C., Miller, B. 1999: A System for Relational Reasoning in the Prefrontal Cortex. ***Psychological Science,*** **10, 119-125.**

Whitfield-Gabrieli, S., Moran, J. M., Nieto-Casta√ ±√ ≥N, A., Triantafyllou, C., Saxe, R. and Gabrieli, J. D. E. 2011: Associations and dissociations between default and self-reference networks in the human brain. ***Neuroimage,*** **55, 225-232.**

Whitfield-Gabrieli, S., Thermenos, H. W., Milanovic, S., Tsuang, M. T., Faraone, S. V., Mccarley, R. W., Shenton, M. E., Green, A. I., Nieto-Castanon, A. and Laviolette, P. 2009: Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. ***Proceedings of the National Academy of Sciences,*** **106, 1279.**

Williams, G. and Goldman-Rakic, P. 1995: Modulation of memory fields by dopamine Dl receptors in prefrontal cortex.

Wolfe, G. and Mckenzie, K. 1994: Capgras, Frégoli and Cotard's syndromes and Koro in folie a deuz. ***British Journal of Psychiatry,*** **165, 842.**

Wolpert, D. and Ghahramani, Z. 2000: Computational prinicples of movement. ***Neuroscience. Nature Neuroscience,*** **supplement 3, 1212-1217.**

Woodward, J. Mental causation and neural mechanisms. ***Being Reduced: New Essays on Reduction, Explanation and Causation*****.**

Woodward, J. 2008: Causation and manipulability. ***Stanford Encyclopedia of Philosophy*****.**

Wright, S., Young, A. W. and Hellawell, D. J. 1993: Frégoli delusion and erotomania. ***Journal of Neurology, Neurosurgery, and Psychiatry,*** **56, 322-323.**

Young, A. and De Pauw, K. 2003: One stage is not enough. ***Philosophy, Psychiatry, & Psychology,*** **9, 55-59.**

Young, A. W. and Burton, A. M. 1999: Simulating face recognition: Implications for modelling cognition. ***Cognitive Neuropsychology,*** **16, 1-48.**

Young, A. W. and Leafhead, K. M. 1996: Betwixt life and death: Case studies of the Cotard delusions. In Halligan, P. W. and Marshall, J. C. (eds) ***Method in Madness: Case studies in cognitive neuropsychiatry.*** **Hove, East Sussex: Psychology Press.**

Young, A. W., Leafhead, K. M. and Szulecka, T. K. 1994: The Capgras and Cotard delusions. ***Psychopathology,*** **27, 226-231.**

Young, A. W., Reid, I., Wright, S. and Hellawell, D. J. 1993: Face-processing impairments and the Capgras delusion. ***British Journal of Psychiatry,*** **162, 695-698.**

Young, G. 2008: Capgras delusion: An interactionist model. ***Consciousness and cognition,*** **17, 863-876.**

Young, G. 2011: Beliefs, experiences and misplaced being: an interactionist account of delusional misidentification. ***Phenomenology and the Cognitive Sciences,*** **10, 195-215.**

FIGURES: (copied from relevant papers as a guide for readers. final will have proper artwork)

Forward model. from Wolpert Doya Kawato 2003

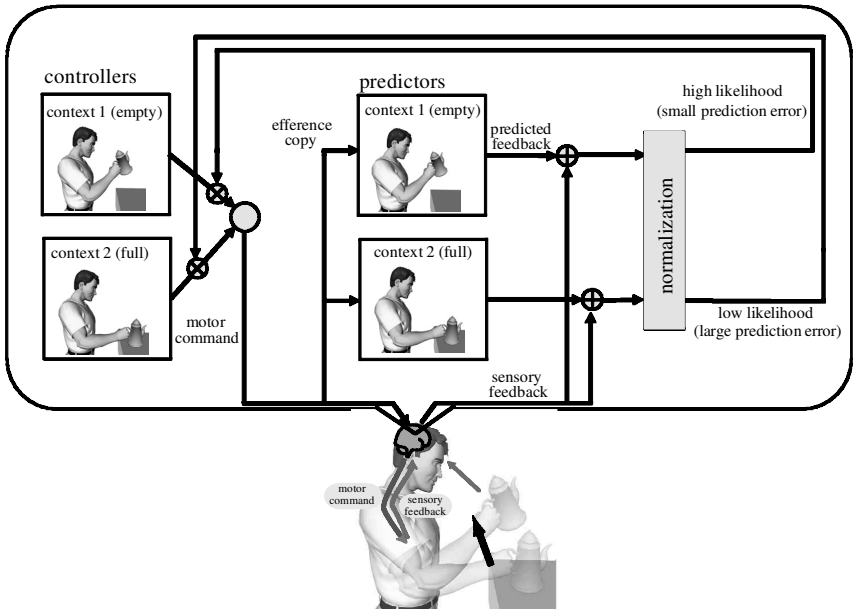

Figure 2. The MOSAIC architecture. A schematic of context estimation with just two contexts: that a teapot is empty or ful In this highly simplified example, a module consists of a controller–predictor pair. In this case two controller–predictor pairs exist: one tuned for a full teapot and one for an empty teapot. The outputs of the controllers are weighted by the likelihood that each is appropriate, to determine the final motor command. When this motor command is generated, an efference copy of the motor command is used to simulate, using the two predictors, the sensory consequences under the two possible contexts. The predictions based on an empty teapot suggest that lift-off will take place early compared with a full teapot and that the lift will be higher. These predictions are compared with actual feedback and the errors are normalized to turn them into likelihood or responsibilities. As the teapot is, in fact, empty the sensory feedback matches the predictions of the empty teapot context. This leads to a high likelihood for the empty teapot and a low likelihood of the full teapot. These responsibilities are used to adjust the weightings of the controllers so as to generate motor commands appropriate for an empty teapot. In addition, the responsibilities are used to gate the learning of the predictors and controllers (not shown).

Forward model: from Frith Blakemore Wolpert 2000

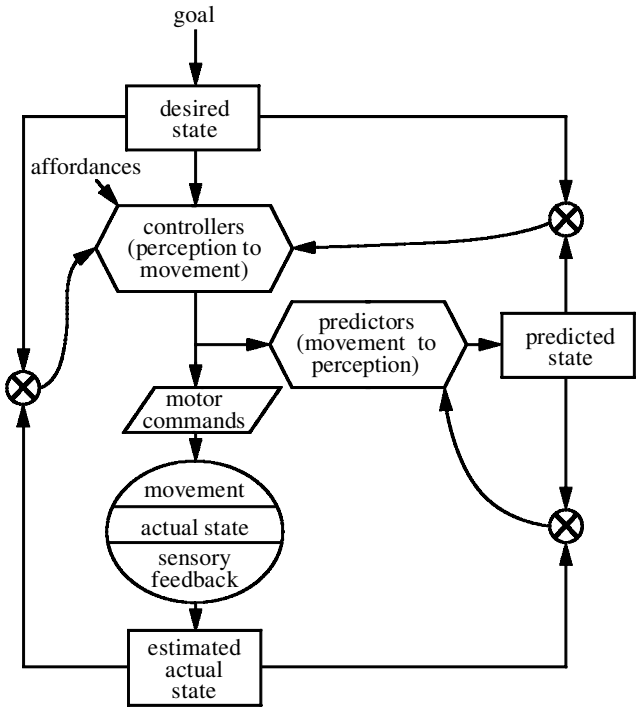

Figure 1. The basic components of a motor control system based upon engineering principles.

(iii) Predicted next state of the system. This representation provides an estimate of the future state of the system derived from the predictors.

(iv) Motor commands. These are derived from the controllers and are fine-tuned by sensory information (affordances) about the current state of the world (e.g. visual information about the position and shape of the object that is to be grasped).

(v) Sensory feedback. This is the consequence of the action performed, plus any environmental events.

Face Recognition, from Burton 2003.